Supplementary material for the article ”Simultaneous estimation of transcript abundances and transcript specific fragment distributions of RNA-Seq data with the Mix^2 model”

Andreas Tuerk, Gregor Wiktorin

Lexogen GmbH
Campus Vienna Biocenter 5, 1030 Vienna, Austria

April 2014

Contents

List of Figures

List of Tables

1 EM update formulas for the Mix^2 model

The Expectation Maximiziation (EM) algorithm [1] increases the likelihood $L(R|\theta)$ of a data set R under a model $p(R|\theta)$ by maximizing, or more generally increasing, the auxiliary function

$$Q(\theta'|\theta) = E_{Z|R,\theta}(\log p(R,z|\theta')) \tag{1}$$

Here, θ is the current parameter set of the model $p(R|\theta)$ and θ' is the new parameter set that needs to be optimized. In addition, $Z = (z_r)_{r\in R}$ is a sequence of random hidden variables z_r and, hence, the expression on the right hand side of (1) is the expected value of $\log p(R,z|\theta')$, where z is one realization of Z, with respect to the random variable Z given R and θ.

The hidden variables in the Mix^2 model are the transcript variable, $t = i$, and the block variable, $b = j$. Since these variables are independent of each other for different fragments r, for the Mix^2 model (1) reduces to

$$Q(\theta'|\theta) = \sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{r\in R} p(t=i,b=j|r,\theta)\log p(r,t=i,b=j|\theta') \tag{2}$$

A necessary condition for the maximization of $Q(\theta'|\theta)$ is that the gradient of $Q(\theta'|\theta)$ equals zero, i.e.

$$\frac{\partial}{\partial\theta'}Q(\theta'|\theta) = 0 \tag{3}$$

If $\theta'_i, \theta'_j \in \theta'$ are parameters in the parameter set θ' of the Mix^2 model, which only depend on $t = i$ and $b = j$ respectively, then (3) implies

$$\frac{\partial}{\partial\theta'_i}Q(\theta'|\theta) = \sum_{j=1}^{M}\sum_{r\in R} p(t=i,b=j|r,\theta)\frac{\partial}{\partial\theta'_i}\log p(r,t=i,b=j|\theta') \tag{4}$$

and

$$\frac{\partial}{\partial\theta'_j}Q(\theta'|\theta) = \sum_{i=1}^{N}\sum_{r\in R} p(t=i,b=j|r,\theta)\frac{\partial}{\partial\theta'_j}\log p(r,t=i,b=j|\theta') \tag{5}$$

1.1 EM update formulas for the continuous Mix^2 model

For the continuous Mix^2 model the conditional probability $p(r|t=i,b=j)$ is given as follows

$$p(r|t=i,b=j) = \frac{1}{\lambda_i\sigma_j\sqrt{2\pi}}e^{-\frac{(r-\lambda_i\mu_j-\nu_i)^2}{2(\lambda_i\sigma_j)^2}} \tag{6}$$

Since the parameters in (6) depend only on either $t = i$ or $b = j$, equations (4) and (5) are necessary conditions for the maximization of the auxiliary function with respect to the parameters in (6). For μ_j, for instance, this yields

$$\sum_{i=1}^{N}\sum_{r\in R} p^{(n)}(t=i,b=j|r)\frac{\left(r-\mu_j^{(n+1)}\lambda_i^{(n)}-\nu_i^{(n)}\right)}{\left(\lambda_i^{(n)}\right)^2\left(\sigma_j^{(n)}\right)^2} = 0 \tag{7}$$

which, after rearrangement, leads to the following update formula for μ_j

$$\mu_j^{(n+1)} = \frac{\sum_{i=1}^{N}\sum_{r\in R} p^{(n)}(t=i,b=j|r)\frac{r-\nu_i^{(n)}}{\lambda_i^{(n)}}}{\sum_{i=1}^{N}\sum_{r\in R} p^{(n)}(t=i,b=j|r)} \tag{8}$$

The expression $\left(r-\nu_i^{(n)}\right)/\lambda_i^{(n)}$ on the right hand side of (8) can be interpreted as a mapping of fragment r onto a common scale which is independent of the length of transcript $t = i$. Thus, the right hand side of (8) is the average of the fragments r on the common scale weighted by the probability of $t = i$.

In a similar manner, the update formula for σ_j can be found to be

$$\left(\sigma_j^{(n+1)}\right)^2 = \frac{\sum_{i=1}^{N}\sum_{r\in R} p^{(n)}(t=i,b=j|r)\left(\frac{r-\nu_i^{(n)}}{\lambda_i^{(n)}}-\mu_j^{(n)}\right)^2}{\sum_{i=1}^{N}\sum_{r\in R} p^{(n)}(t=i,b=j|r)} \tag{9}$$

As the update formula for μ_j, equation (9) has an intuitive interpretation as the variance of the fragments r on the common scale weighted by $p^{(n)}(t=i,b=j|r)$.

For the shift and scale parameters ν_i and λ_i, which depend only on $t=i$, (4) is a necessary condition for the maximization of the auxiliary function which together with

$$\frac{\partial}{\partial \nu_i}\log p(r|t=i,b=j) = \frac{r-\mu_j\lambda_i-\nu_i}{\lambda_i^2\sigma_i^2} \tag{10}$$

yields the following update formula for ν_i.

$$\nu_i^{(n+1)} = \frac{\sum_{j=1}^{M}\sum_{r\in R} p^{(n)}(t=i,b=j|r)\frac{r-\lambda_i^{(n)}\mu_j^{(n)}}{\left(\sigma_j^{(n)}\right)^2}}{\sum_{j=1}^{M}\sum_{r\in R} p^{(n)}(t=i,b=j|r)\frac{1}{\left(\sigma_j^{(n)}\right)^2}} \tag{11}$$

Likewise (4) together with

$$\frac{\partial}{\partial \lambda_i}\log p(r|t=i,b=j) = -\frac{1}{\lambda_i}+\frac{1}{\sigma_j^2}\left(\frac{(r-\nu_i)^2}{\lambda_i^3}-\frac{\mu_j(r-\nu_i)}{\lambda_i^2}\right) \tag{12}$$

leads to the following quadratic equation for $\lambda_i^{(n+1)}$.

$$\begin{aligned} & \left(\lambda_i^{(n+1)}\right)^2\sum_{j=1}^{M}\sum_{r\in R} p^{(n)}(t=i,b=j|r) \\ + \quad & \lambda_i^{(n+1)}\sum_{j=1}^{M}\sum_{r\in R} p^{(n)}(t=i,b=j|r)\frac{\mu_j\left(r-\nu_i^{(n)}\right)}{\left(\sigma_j^{(n)}\right)^2} \\ - \quad & \sum_{j=1}^{M}\sum_{r\in R} p^{(n)}(t=i,b=j|r)\frac{\left(r-\nu_i^{(n)}\right)^2}{\left(\sigma_j^{(n)}\right)^2} \\ = \quad & 0 \end{aligned} \tag{13}$$

For sensible initial values of the parameters equation (13) has one positive and one negative solution. Since λ_i is a scaling factor only the positive solution of (13) is of interest.

1.2 EM update formulas for the Mix^2 model with group tying

If transcripts are separated into different groups the update formulas of the parameters which are tied between transcripts have to be modified. These are the parameters β_j, μ_j and σ_j whose update formulas are presented in this section. In the following, each transcript $t=i$ has an associated group $g=k$ which can be retrieved via the function $G(i)=k$.

Using the Lagrange method to enforce the constraint

$$\sum_{j=1}^{M}\beta_{k,j} = 1 \tag{14}$$

and taking the derivative with respect to $\beta_{k,j}$ leads to

$$\sum_{i:G(i)=k}\sum_{r\in R} p(t=i,b=j|r)+\beta_{k,j}\lambda = 0 \tag{15}$$

which after some rearrangement results in

$$\beta_{k,j}^{(n+1)} = \frac{\sum_r p^{(n)}\left(g=k, b=j|r\right)}{\sum_r p^{(n)}\left(g=k|r\right)} \tag{16}$$

where

$$p(g=k, b=j|r) = \sum_{i:G(i)=k} p(t=i, b=j|r) \tag{17}$$

For the remaining parameters which are tied between transcripts, i.e. μ_j and σ_j, the update formulas (8) and (9) need to be modified by replacing the sum over the complete set of transcripts $t = 1, \ldots, N$ by the sum over the transcripts within the group, i.e.

$$\mu_{k,j}^{(n+1)} = \frac{\sum_{i:G(i)=k}\sum_{r\in R} p^{(n)}(t=i, b=j|r)\frac{r-\nu_i^{(n)}}{\lambda_i^{(n)}}}{\sum_{r\in R} p^{(n)}(g=k, b=j|r)} \tag{18}$$

$$\left(\sigma_j^{(n+1)}\right)^2 = \frac{\sum_{i:G(i)=k}\sum_{r\in R} p^{(n)}(t=i, b=j|r)\left(\frac{r-\nu_i^{(n)}}{\lambda_i^{(n)}} - \mu_j^{(n)}\right)^2}{\sum_{r\in R} p^{(n)}(g=k, b=j|r)} \tag{19}$$

2 Combining the Mix^2 model with other bias models

The Mix^2 parameter tying discussed in the main paper implements a model for the positional fragmentation bias, i.e. the bias related to the fragment start within a transcript. Other kinds of bias, like the sequence specific bias, might be described with other models such as the variable length hidden Markov model (VLMM) in [2]. Typically, models for non-positional bias compare the observed frequency of nucleotide sequences to their frequency under the null-hypothesis of unbiased data. This comparison can be used to derive a probability distribution $p(m = c|r)$ over the multiplicity $m = c$ of fragment r in the absence of any bias, given that a single copy of r is observed in the biased data. The distribution $p(m = c|r)$ can then be used to computationally remove the non-positional bias from the data by weighting each fragment r in the EM update formulas of the Mix^2 model by the expected multiplicity of r. For the EM update formula of the abundances α_i, for instance, this leads to

$$\alpha_i^{(n+1)} = \frac{1}{|R|} \sum_{r \in R} \sum_{c} c p(m = c|r) p^{(n)}(t = i|r) \tag{20}$$

It should be noted that the distribution $p(m = c|r)$ cannot be estimated by maximizing the likelihood of the data set R under the Mix^2 model, as this would lead to infinite expected values of $p(m = c|r)$. For the combination of the Mix^2 model with other bias models, as discussed in this section, it is therefore important to estimate $p(m = c|r)$ outside the maximum likelihood framework of the Mix^2 model.

3 Properties of the test data

This section collects some of the properties of the test data that were used in the experiments in the main paper.

Table 1 to Table 7 contain the names of the transcripts together with their length, Ensembl id and transcript id. The latter was used in the experiments to improve readability and appears in the graphics in Section 4 and Section 5.

Figure 1 to Figure 4 show examples for the coverage and distribution of 1000 fragments sampled for the HAUS5 gene with 5' bias, Figure 1, with 3' bias, Figure 2, with 5'+3' bias, Figure 3, and with Cufflinks bias, Figure 4. These figures show screenshots of the IGV browser [3], where the upper track of the IGV session visualizes the coverage of the sampled fragments while the middle track shows the distribution of the sampled reads and their splicing. The track at the bottom of the figures shows the annotation of the HAUS5 gene, which consists of 10 transcripts. As mentioned in the main paper, comparing Figure 1 and Figure 2 with Figure 3 and Figure 4 suggests that the latter exhibit a greater similarity to each other than any other pair of biases, which explains the good performance of Cufflinks 2.2.0 on data with 5'+3' bias.

name	Ensembl id	length	transcript id
TESK2-201	ENST00000341771	2989	1
TESK2-002	ENST00000372084	2600	2
TESK2-001	ENST00000372086	3074	3
TESK2-202	ENST00000451835	950	4
TESK2-003	ENST00000486676	3019	5
TESK2-004	ENST00000493974	428	6
TESK2-203	ENST00000538496	2327	7

Table 1: Names, ids and lengths of transcripts in TESK2 gene.

name	Ensembl id	length	transcript id
KLK5-004	ENST00000593428	1365	1
KLK5-001	ENST00000336334	1563	2
KLK5-005	ENST00000391809	1405	3
KLK5-002	ENST00000595585	1301	4
KLK5-006	ENST00000594846	672	5

Table 2: Names, ids and lengths of transcripts in KLK5 gene.

name	Ensembl id	length	transcript id
LDHD-001	ENST00000300051	2067	1
LDHD-002	ENST00000450168	1966	2
LDHD-004	ENST00000568164	740	3
LDHD-003	ENST00000569876	701	4

Table 3: Names, ids and lengths of transcripts in LDHD gene.

name	Ensembl id	length	transcript id
LGALS17A-001	ENST00000412609	2562	1
LGALS17A-201	ENST00000455832	2584	2
LGALS17A-002	ENST00000458539	1807	3
LGALS17A-006	ENST00000598164	406	4
LGALS17A-004	ENST00000598304	2018	5
LGALS17A-005	ENST00000598736	1111	6

Table 4: Names, ids and lengths of transcripts in LGALS17A gene.

name	Ensembl id	length	transcript id
DAPK3-001	ENST00000545797	2257	1
DAPK3-004	ENST00000595279	2058	2
DAPK3-005	ENST00000596311	671	3
DAPK3-006	ENST00000601824	599	4
DAPK3-007	ENST00000593844	626	5
DAPK3-008	ENST00000594894	617	6
DAPK3-201	ENST00000301264	2105	7

Table 5: Names, ids and lengths of transcripts in DAPK3 gene.

name	Ensembl id	length	transcript id
HAUS5-002	ENST00000428854	4387	1
HAUS5-003	ENST00000424522	4462	2
HAUS5-004	ENST00000430749	526	3
HAUS5-005	ENST00000203166	4283	4
HAUS5-006	ENST00000587439	2758	5
HAUS5-007	ENST00000588570	553	6
HAUS5-008	ENST00000592291	526	7
HAUS5-009	ENST00000585968	2762	8
HAUS5-010	ENST00000590994	883	9
HAUS5-201	ENST00000379045	2477	10

Table 6: Names, ids and lengths of transcripts in HAUS5 gene.

name	Ensembl id	length	transcript id
USF2-001	ENST00000343550	1523	1
USF2-002	ENST00000379134	648	2
USF2-003	ENST00000594264	2742	3
USF2-004	ENST00000595068	1612	4
USF2-005	ENST00000594064	1179	5
USF2-006	ENST00000596380	550	6
USF2-007	ENST00000602164	582	7
USF2-009	ENST00000598058	510	8
USF2-008	ENST00000599625	446	9
USF2-010	ENST00000599471	1173	10
USF2-012	ENST00000593708	474	11
USF2-013	ENST00000222305	1634	12
USF2-015	ENST00000597671	464	13
USF2-014	ENST00000600341	1392	14
USF2-016	ENST00000600898	848	15

Table 7: Names, ids and lengths of transcripts in USF2 gene.

transcript id	abundance
1	0.026894
2	0.168888
3	0.152779
4	0.074446
5	0.139321
6	0.139827
7	0.088364
8	0.016049
9	0.076817
10	0.116615

Table 8: Abundances of the HAUS5 transcripts in the examples in Figure 1 to Figure 4.

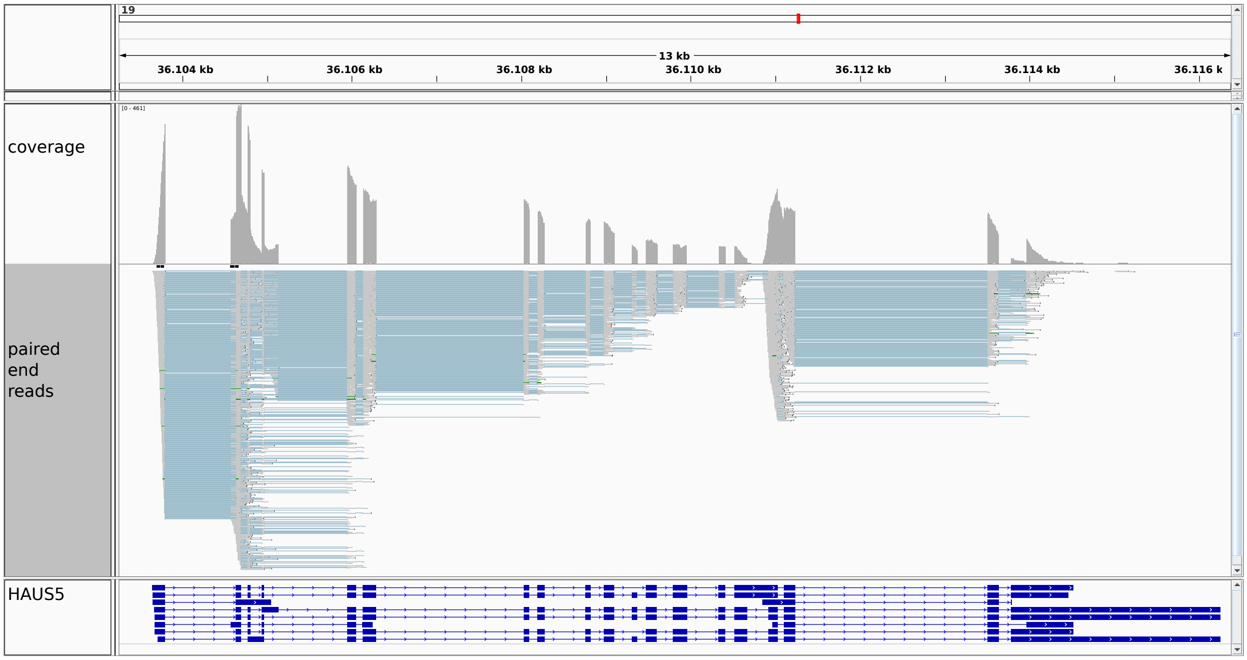

Figure 1: Coverage and distribution of 1000 sampled paired end reads sampled from the HAUS5 gene with a 5' bias. This figure shows a snapshot of the IGV browser. The top track in this figure shows the coverage of the data, while the middle track shows the individual paired end reads. The bottom track shows the Ensembl annotation of the HAUS5 gene.

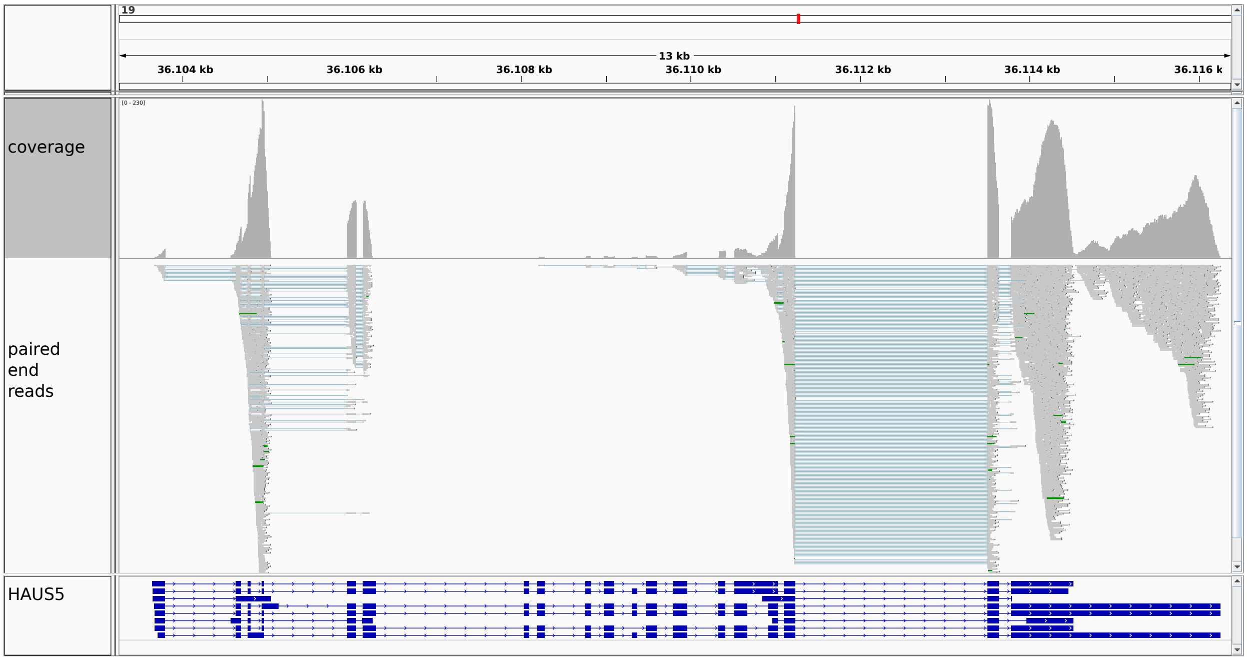

Figure 2: Coverage and distribution of 1000 sampled paired end reads sampled from the HAUS5 gene with a 3' bias. This figure shows a snapshot of the IGV browser. The top track in this figure shows the coverage of the data, while the middle track shows the individual paired end reads. The bottom track shows the Ensembl annotation of the HAUS5 gene.

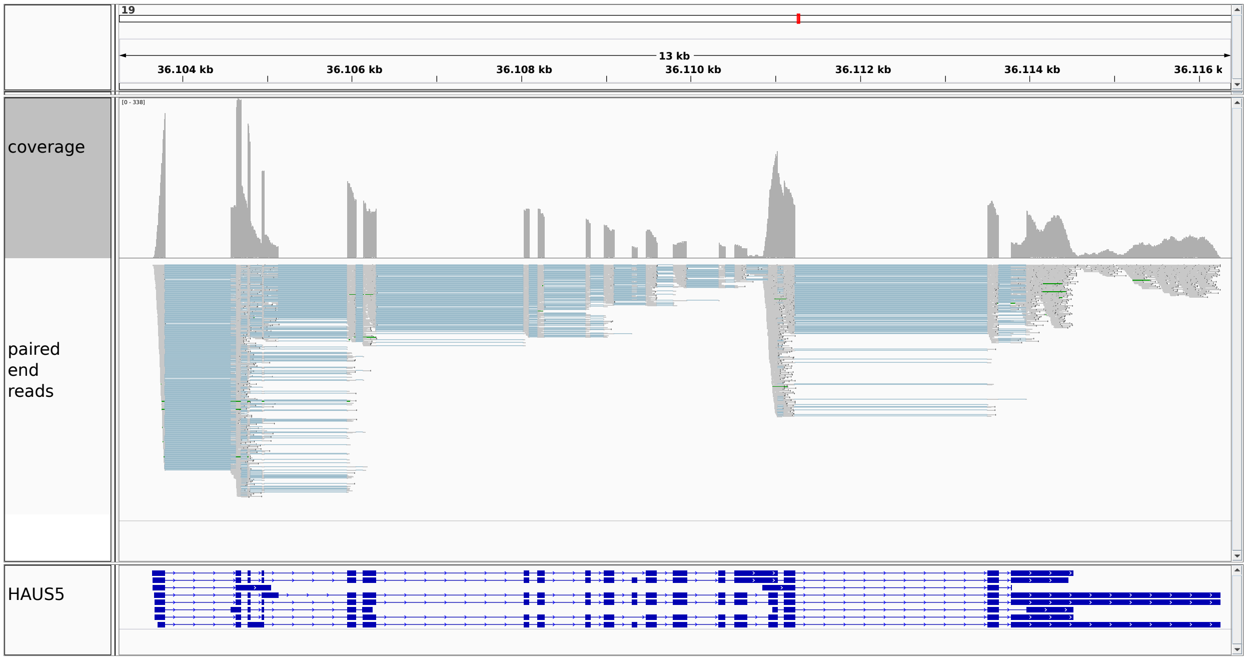

Figure 3: Coverage and distribution of 1000 sampled paired end reads sampled from the HAUS5 gene with a 5'+3' bias. This figure shows a snapshot of the IGV browser. The top track in this figure shows the coverage of the data, while the middle track shows the individual paired end reads. The bottom track shows the Ensembl annotation of the HAUS5 gene.

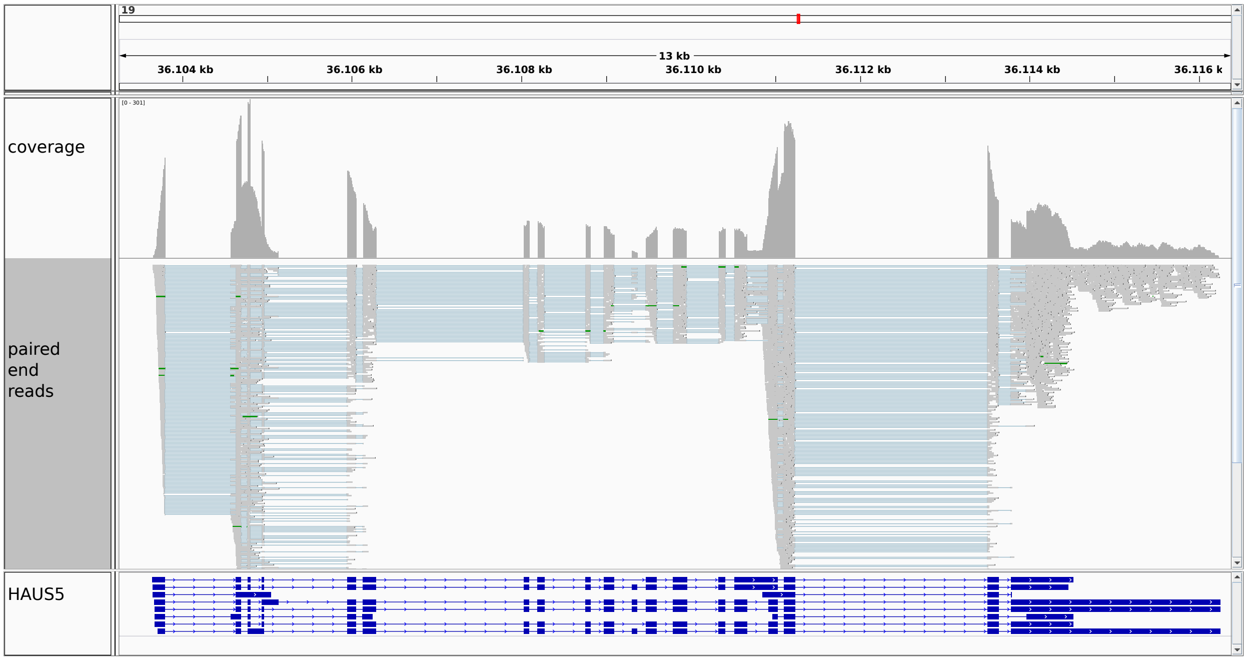

Figure 4: Coverage and distribution of 1000 sampled paired end reads sampled from the HAUS5 gene with a Cufflinks bias. This figure shows a snapshot of the IGV browser. The top track in this figure shows the coverage of the data, while the middle track shows the individual paired end reads. The bottom track shows the Ensembl annotation of the HAUS5 gene.

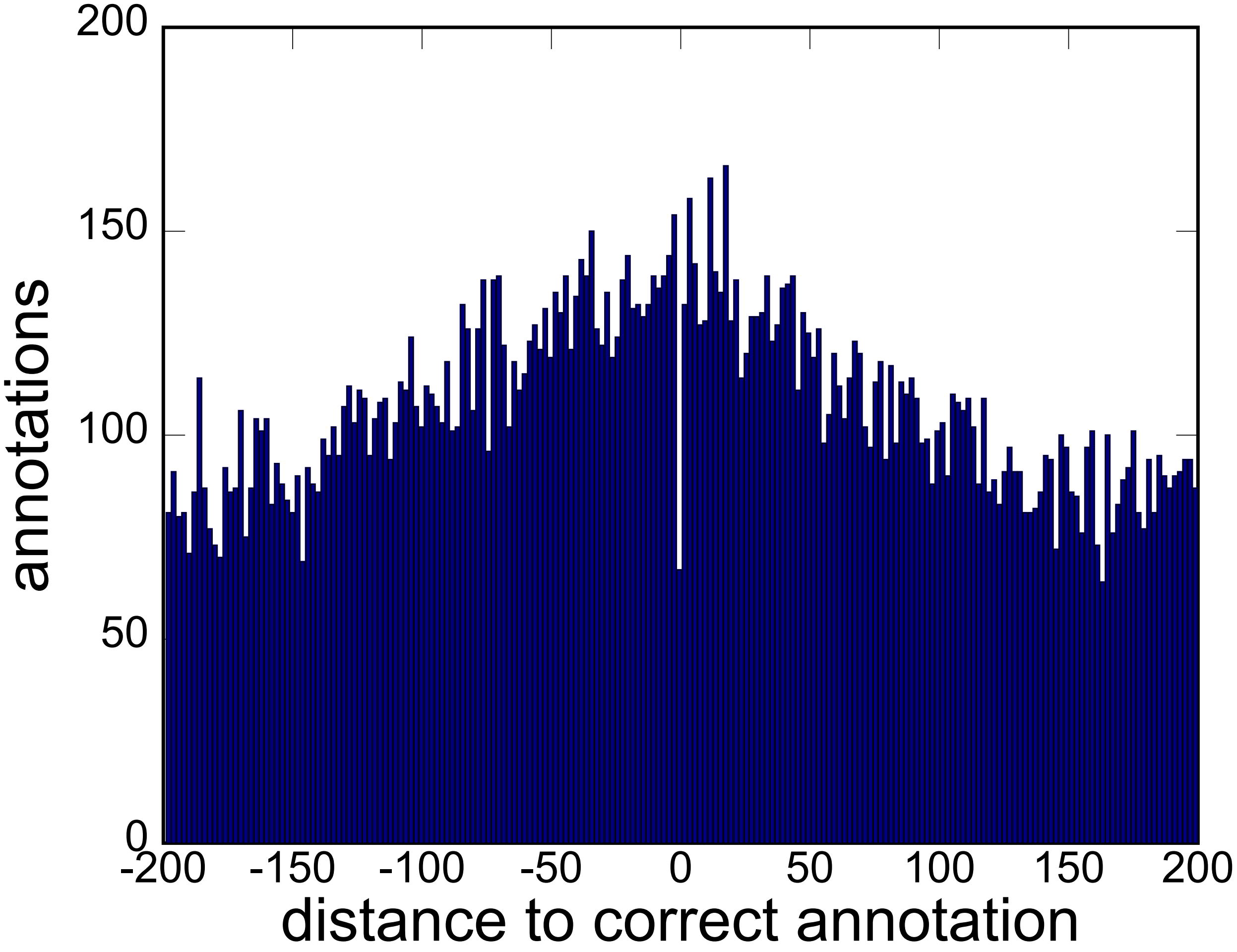

Figure 5: Histogram of the difference between correct and incorrect annotation in the experiments in Section 3.3 of the main paper.

4 Cufflinks, PennSeq and the Mix^2 model on correct annotations

This section contains the detailed results of the experiments on correct annotations comparing Cufflinks 2.2.0, PennSeq and the 2p Mix^2 model with and without group tying.

Figure 6 shows the average L_1 distance between true and estimated abundances for the 4 different biases, which, together with the standard deviations, are given in Table 9 to Table 12. Figure 5 in the main paper is a summary of Figure 6 in the supplement.

For each gene and each type of bias, boxplots of the difference as well as the L_1 distance between true and estimated abundances are given in Figure 7 to Figure 34. These boxplots summarize the results for Cufflinks 2.2.0, PennSeq and the 2p Mix^2 model with group tying for the 200 abundance sets that were sampled, according to the Dirichlet distribution, for each gene and bias.

The figures in this section show that, with the exception of the Cufflinks bias, the 2p Mix^2 model consistently outperforms both Cufflinks 2.2.0 and PennSeq. Only for data sampled from the Cufflinks fragment distribution does Cufflinks 2.2.0 yield a slightly more accurate result than the 2p Mix^2 model with group tying.

Comparing the results for the 2p Mix^2 model with and without group tying shows that group tying yields improved abundance estimates for all bias types apart from the 5' bias. This is due to the fact that for data with 5' bias the transcript length dependent variability of the fragment start distributions is low, as can be seen from Figure 3(b) in the main paper. Thus the introduction of two groups to separate long and short transcripts results in an unnessary increase in model parameters, which leads to slight overfitting.

It is interesting to note that the number of transcripts within a gene is not the only factor influencing the difficulty of abundance estimation. The USF2 gene containing 15 transcripts, for instance, does not yield the highest average L$_1$ distance for any bias or any estimation method. On the other hand, the TESK2 gene, which contains only 7 transcripts, is always among the most difficult genes for any bias and estimation method. This is most likely due to the fact that the TESK2 gene has a single splicing variant, as can be seen in Table 1 in the main paper, and that therefore the TESK2 transcripts are more confusable than that of other genes.

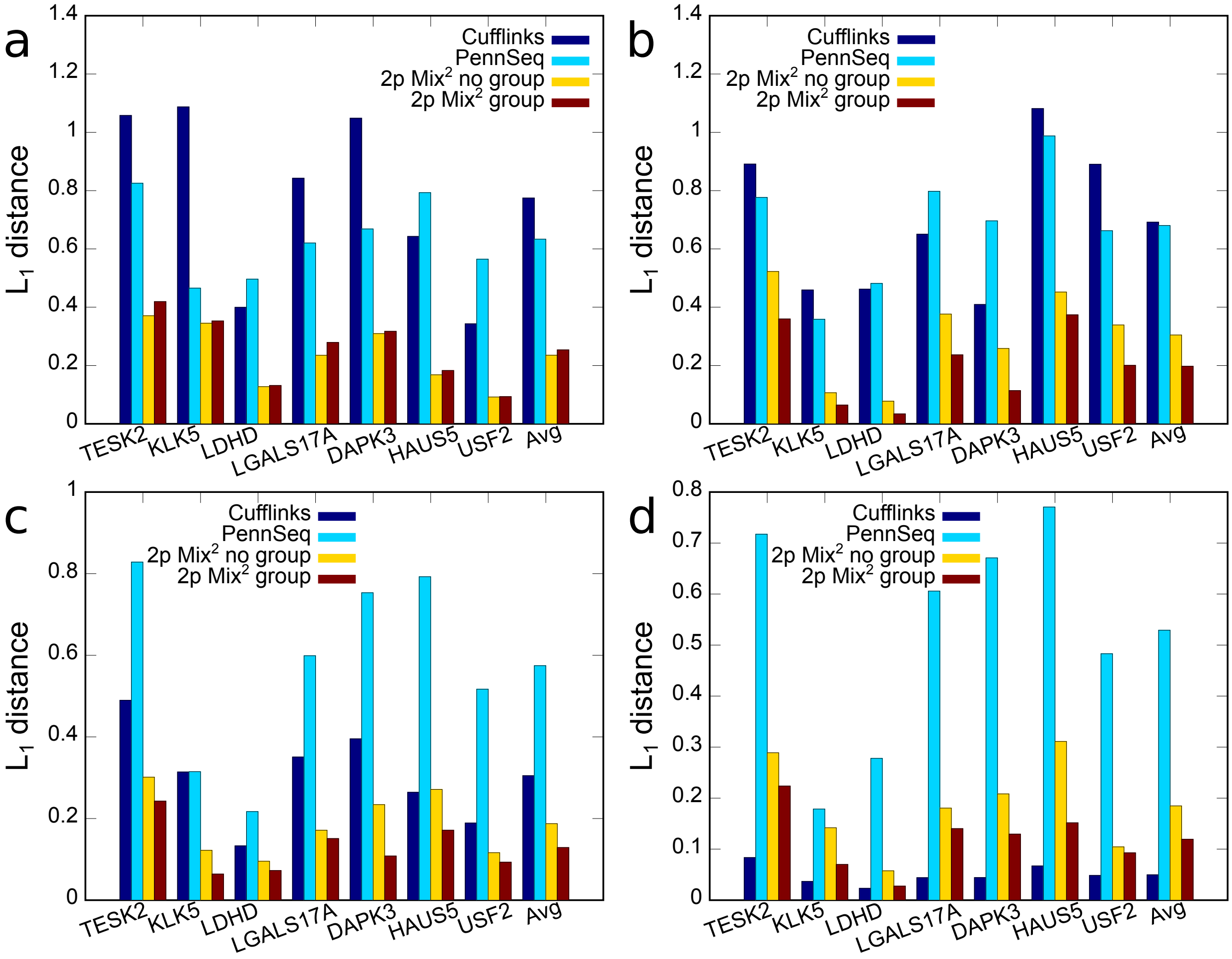

Figure 6: Average L_1 distance between true and estimated abundance with correct annotations and data with 5' bias (a), 3' bias (b), 5'+3' bias (c) and Cufflinks bias (d). The corresponding numbers, including the standard deviation of the L_1 distance can be found in Table 9 for data with 5' bias, in Table 10 for data with 3' bias, in Table 11 for data with 5'+3' bias and in Table 12 for data with Cufflinks bias.

Gene	Cufflinks		PennSeq		2p Mix^2 no group		2p Mix^2 group	
	mean	std	mean	std	mean	std	mean	std
TESK2	1.05782	0.32478	0.82548	0.19293	0.37099	0.17225	0.41922	0.15830
KLK5	1.08703	0.35782	0.46551	0.20802	0.34509	0.18183	0.35278	0.18272
LDHD	0.39955	0.16921	0.49632	0.14201	0.12765	0.13576	0.13140	0.13925
LGALS17A	0.84243	0.30474	0.62040	0.14089	0.23525	0.12746	0.27919	0.15331
DAPK3	1.04828	0.31805	0.66836	0.21532	0.30965	0.13476	0.31732	0.13810
HAUS5	0.64286	0.24876	0.79311	0.13297	0.16819	0.10399	0.18278	0.10918
USF2	0.34316	0.10607	0.56483	0.10430	0.09190	0.04017	0.09311	0.04190

Table 9: Mean and standard deviation of L_1 distance between true and estimated abundance with correct annotations and data with 5' bias. The means are visualized in Table 6(a)

Gene	Cufflinks		PennSeq		2p Mix^2 no group		2p Mix^2 group	
	mean	std	mean	std	mean	std	mean	std
TESK2	0.89095	0.24114	0.77651	0.16560	0.52250	0.19795	0.35987	0.12964
KLK5	0.45914	0.24514	0.35856	0.20758	0.10657	0.06377	0.06420	0.02918
LDHD	0.46210	0.13652	0.48159	0.15413	0.07765	0.04051	0.03398	0.02019
LGALS17A	0.65048	0.27459	0.79760	0.26320	0.37661	0.23302	0.23656	0.16732
DAPK3	0.40938	0.17707	0.69613	0.20791	0.25849	0.17339	0.11380	0.07767
HAUS5	1.08137	0.27017	0.98766	0.24541	0.45199	0.12685	0.37383	0.14916
USF2	0.89008	0.17854	0.66241	0.11328	0.33903	0.10192	0.20046	0.06857

Table 10: Mean and standard deviation of L_1 distance between true and estimated abundance with correct annotations and data with 3' bias. The means are visualized in Table 6(b)

Gene	Cufflinks		PennSeq		2p Mix^2 no group		2p Mix^2 group	
	mean	std	mean	std	mean	std	mean	std
TESK2	0.48968	0.18135	0.82845	0.17256	0.30142	0.14103	0.24264	0.11572
KLK5	0.31405	0.09807	0.31484	0.12394	0.12214	0.05069	0.06401	0.02386
LDHD	0.13293	0.05377	0.21700	0.07815	0.09543	0.05815	0.07260	0.04942
LGALS17A	0.35061	0.11197	0.59884	0.19133	0.17136	0.05741	0.15100	0.05008
DAPK3	0.39526	0.13649	0.75324	0.16160	0.23409	0.14566	0.10823	0.05792
HAUS5	0.26420	0.10346	0.79248	0.14305	0.27124	0.07220	0.17148	0.07982
USF2	0.18880	0.04294	0.51700	0.09784	0.11636	0.03978	0.09305	0.03535

Table 11: Mean and standard deviation of L_1 distance between true and estimated abundance with correct annotations and data with 5'+3' bias. The means are visualized in Table 6(c)

Gene	Cufflinks		PennSeq		2p Mix^2 no group		2p Mix^2 group	
	mean	std	mean	std	mean	std	mean	std
TESK2	0.08347	0.04326	0.71742	0.16563	0.28896	0.14186	0.22374	0.12198
KLK5	0.03677	0.01555	0.17865	0.05697	0.14184	0.07795	0.07008	0.02914
LDHD	0.02322	0.01295	0.27807	0.10801	0.05759	0.02423	0.02743	0.01639
LGALS17A	0.04411	0.02055	0.60577	0.13197	0.18043	0.09125	0.14027	0.06396
DAPK3	0.04419	0.01998	0.67099	0.16583	0.20836	0.13621	0.12948	0.07453
HAUS5	0.06703	0.02955	0.77058	0.16586	0.31114	0.12204	0.15165	0.06422
USF2	0.04858	0.01287	0.48315	0.09558	0.10458	0.02770	0.09266	0.02477

Table 12: Mean and standard deviation of L_1 distance between true and estimated abundance with correct annotations and data with Cufflinks bias. The means are visualized in Table 6(d)

Gene	Cufflinks		PennSeq		2p Mix^2 no group		2p Mix^2 group	
	mean	std	mean	std	mean	std	mean	std
TESK2	0.63048	0.19763	0.78696	0.17418	0.37097	0.16327	0.31137	0.13141
KLK5	0.47425	0.17914	0.32939	0.14913	0.17891	0.09356	0.13777	0.06623
LDHD	0.25445	0.09311	0.36825	0.12057	0.08958	0.06466	0.06635	0.05631
LGALS17A	0.47191	0.17796	0.65565	0.18185	0.24091	0.12729	0.20175	0.10867
DAPK3	0.47428	0.16290	0.69718	0.18766	0.25265	0.14750	0.16721	0.08705
HAUS5	0.51387	0.16298	0.83596	0.17182	0.30064	0.10627	0.21993	0.10060
USF2	0.36766	0.08511	0.55685	0.10275	0.16297	0.05239	0.11982	0.04265

Table 13: Mean and standard deviation of L_1 distance between true and estimated abundance with correct annotations averaged over biases. The means are visualized in Figure 5(a) in the main paper.

bias	Cufflinks		PennSeq		2p Mix^2 no group		2p Mix^2 group	
	mean	std	mean	std	mean	std	mean	std
5' bias	0.77445	0.26135	0.63343	0.16235	0.23553	0.12803	0.25369	0.13182
3' bias	0.69193	0.21760	0.68007	0.19387	0.30469	0.13392	0.19753	0.09168
5'+3' bias	0.30508	0.10401	0.57455	0.13835	0.18743	0.08070	0.12900	0.05888
Cufflinks bias	0.04962	0.02210	0.52923	0.12712	0.18470	0.08875	0.11933	0.05643

Table 14: Mean and standard deviation of L_1 distance between true and estimated abundance with correct annotations averaged over genes. The means are visualized in Figure 5(b) in the main paper.

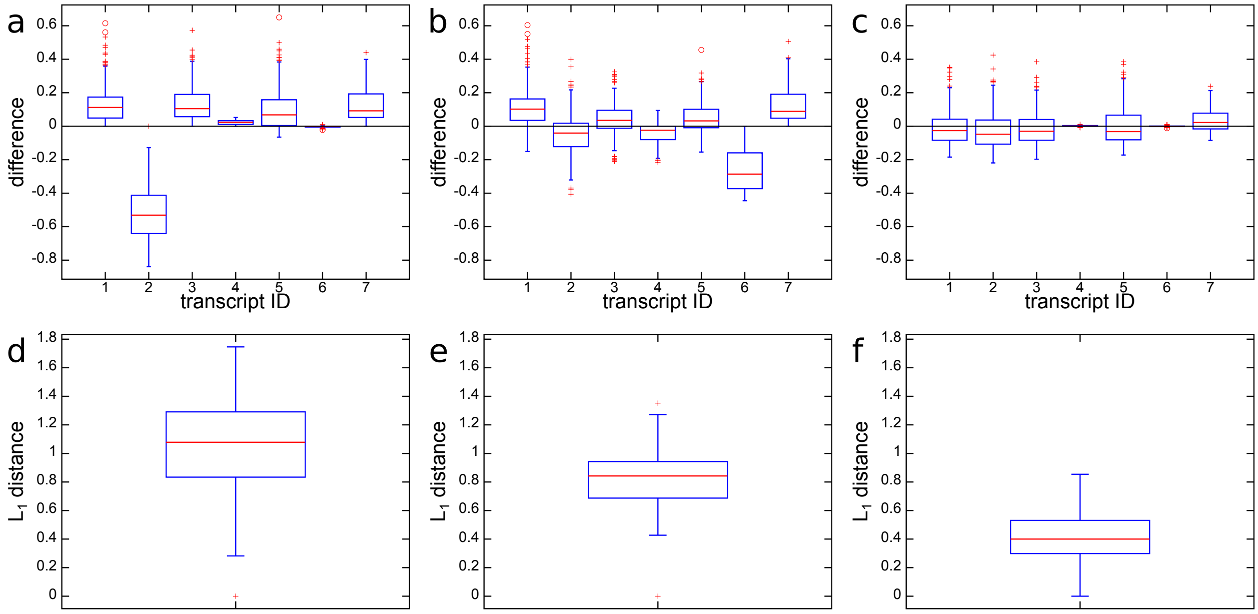

Figure 7: Difference between true and estimated abundances for **data with 5' bias and correct annotations**. Figures (a), (b) and (c) show boxplots for the difference for each of the 7 transcripts in the **TESK2** gene for **Cufflinks 2.2.0** (a), **PennSeq** (b) and the **2p Mix^2 model with group tying** (c). Figures (d), (e) and (f) show boxplots of the L_1 distance for Cufflinks 2.2.0 (d), PennSeq (e) and the 2p Mix^2 model with group tying (f). The mean and the standard deviation of the values in Figures (c) and (d) are given in Table 9.

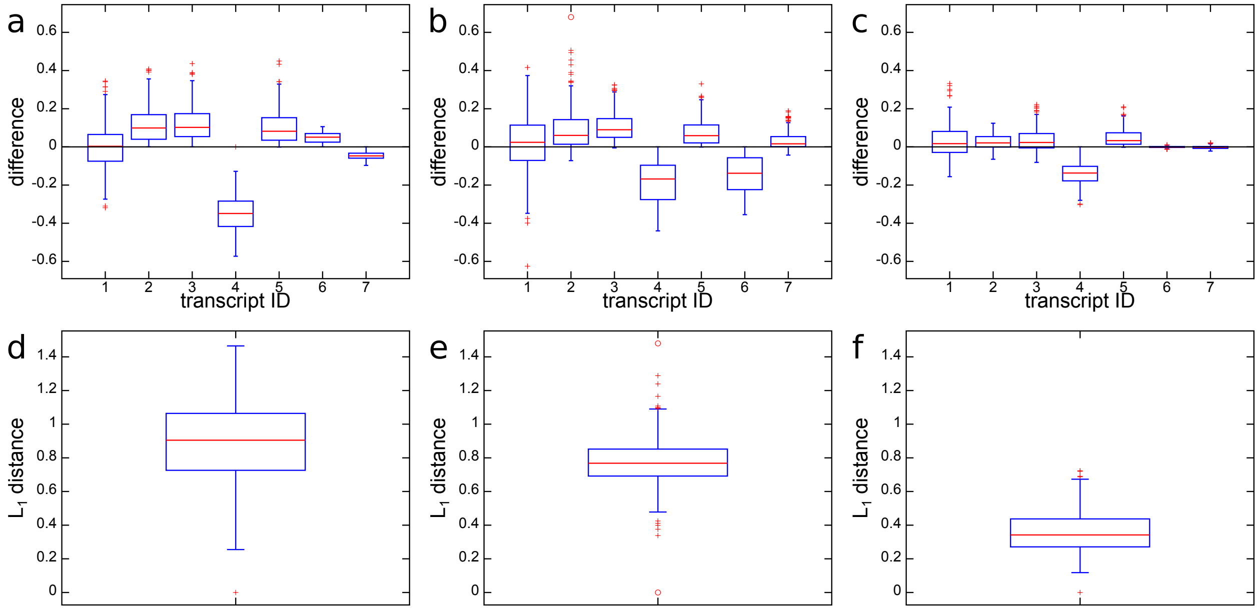

Figure 8: Difference between true and estimated abundances for **data with 3' bias and correct annotations**. Figures (a), (b) and (c) show boxplots for the difference for each of the 7 transcripts in the **TESK2** gene for **Cufflinks 2.2.0** (a), **PennSeq** (b) and the **2p Mix^2 model with group tying** (c). Figures (d), (e) and (f) show boxplots of the L_1 distance for Cufflinks 2.2.0 (d), PennSeq (e) and the 2p Mix^2 model with group tying (f). The mean and the standard deviation of the values in Figures (c) and (d) are given in Table 10.

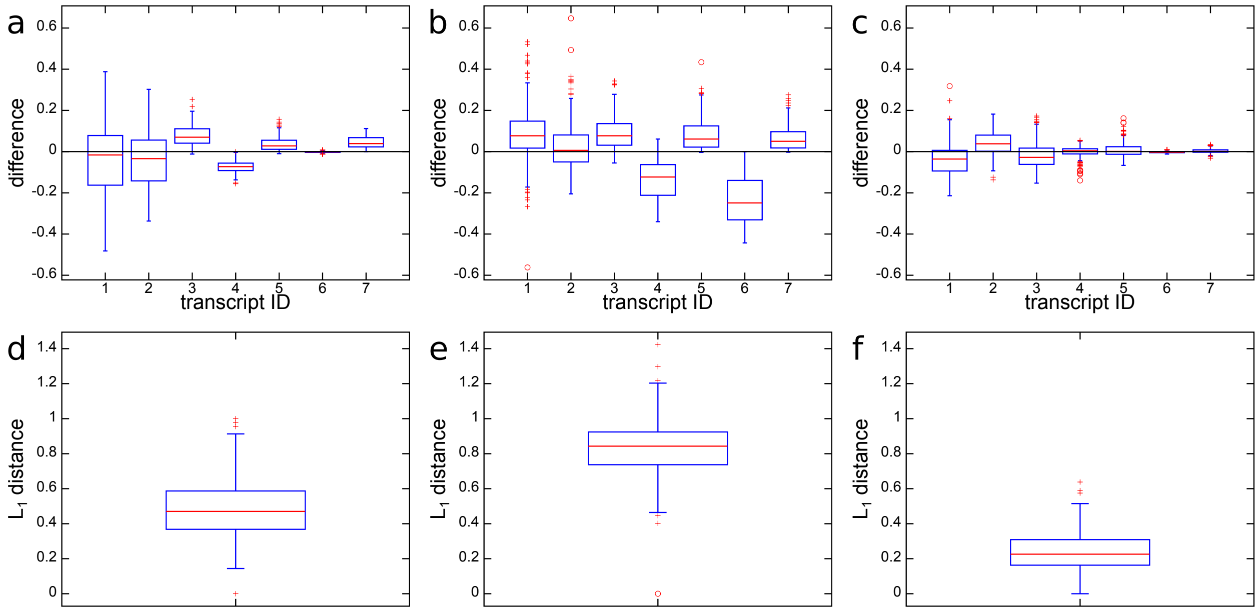

Figure 9: Difference between true and estimated abundances for **data with 5'+3' bias and correct annotations**. Figures (a), (b) and (c) show boxplots for the difference for each of the 7 transcripts in the **TESK2** gene for **Cufflinks 2.2.0** (a), **PennSeq** (b) and the **2p Mix^2 model with group tying** (c). Figures (d), (e) and (f) show boxplots of the L_1 distance for Cufflinks 2.2.0 (d), PennSeq (e) and the 2p Mix^2 model with group tying (f). The mean and the standard deviation of the values in Figures (c) and (d) are given in Table 11.

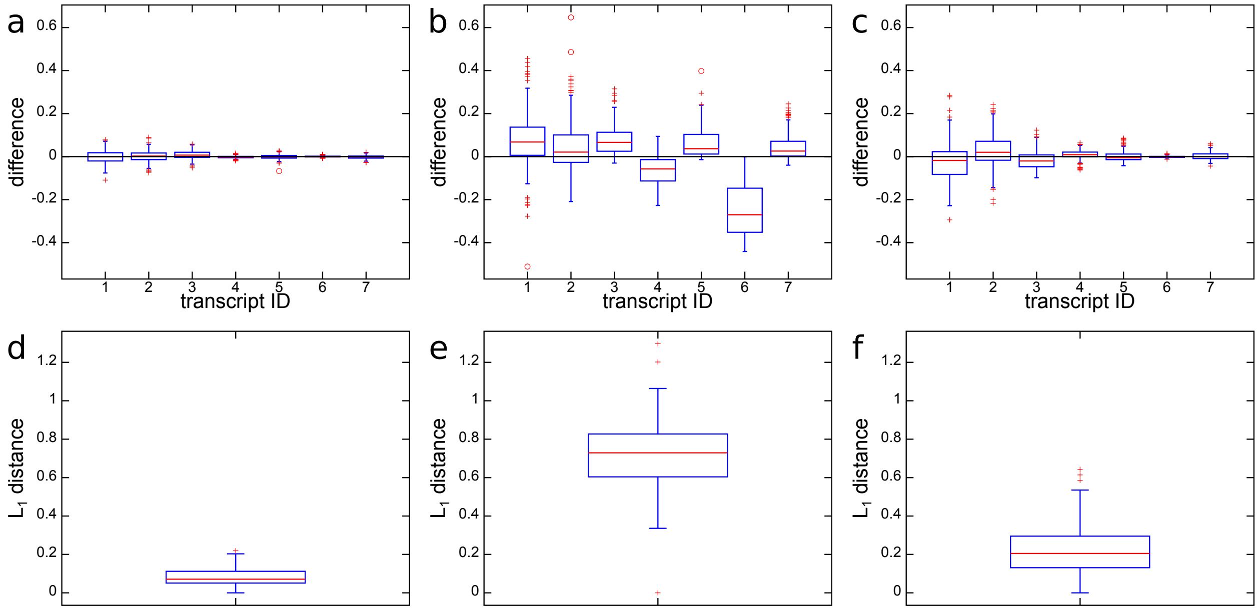

Figure 10: Difference between true and estimated abundances for **data with Cufflinks bias and correct annotations**. Figures (a), (b) and (c) show boxplots for the difference for each of the 7 transcripts in the **TESK2** gene for **Cufflinks 2.2.0** (a), **PennSeq** (b) and the **2p Mix^2 model with group tying** (c). Figures (d), (e) and (f) show boxplots of the L_1 distance for Cufflinks 2.2.0 (d), PennSeq (e) and the 2p Mix^2 model with group tying (f). The mean and the standard deviation of the values in Figures (c) and (d) are given in Table 12.

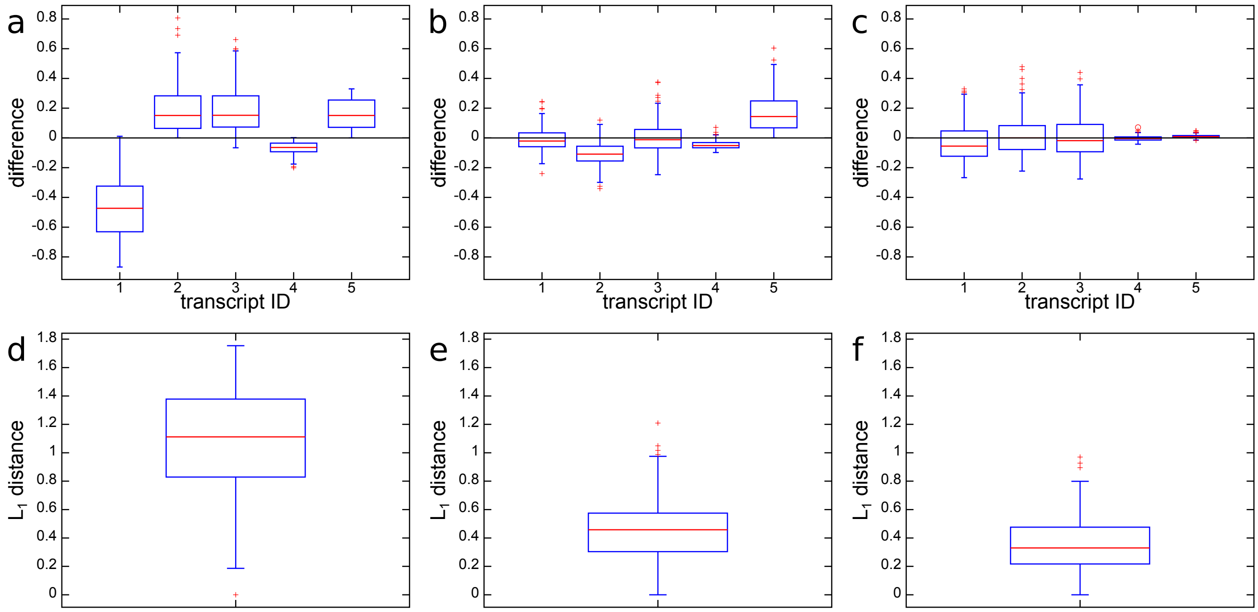

Figure 11: Difference between true and estimated abundances for **data with 5' bias and correct annotations**. Figures (a), (b) and (c) show boxplots for the difference for each of the 5 transcripts in the **KLK5** gene for **Cufflinks 2.2.0** (a), **PennSeq** (b) and the **2p Mix^2 model with group tying** (c). Figures (d), (e) and (f) show boxplots of the L_1 distance for Cufflinks 2.2.0 (d), PennSeq (e) and the 2p Mix^2 model with group tying (f). The mean and the standard deviation of the values in Figures (c) and (d) are given in Table 9.

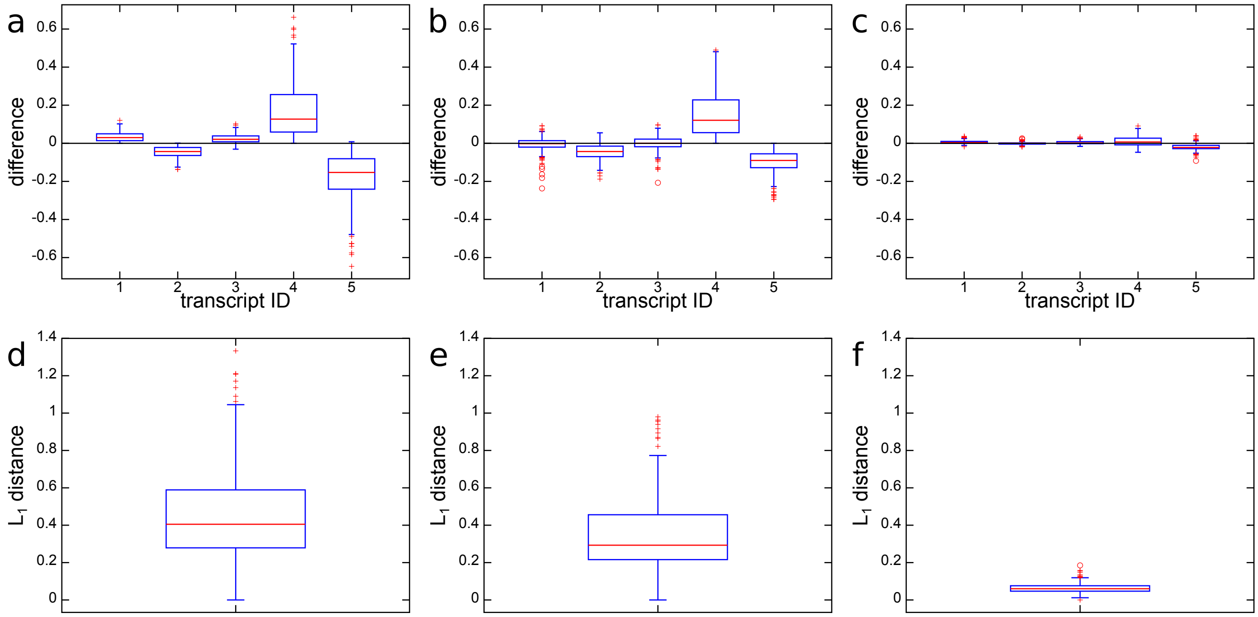

Figure 12: Difference between true and estimated abundances for **data with 3' bias and correct annotations**. Figures (a), (b) and (c) show boxplots for the difference for each of the 5 transcripts in the **KLK5** gene for **Cufflinks 2.2.0** (a), **PennSeq** (b) and the **2p Mix^2 model with group tying** (c). Figures (d), (e) and (f) show boxplots of the L_1 distance for Cufflinks 2.2.0 (d), PennSeq (e) and the 2p Mix^2 model with group tying (f). The mean and the standard deviation of the values in Figures (c) and (d) are given in Table 10.

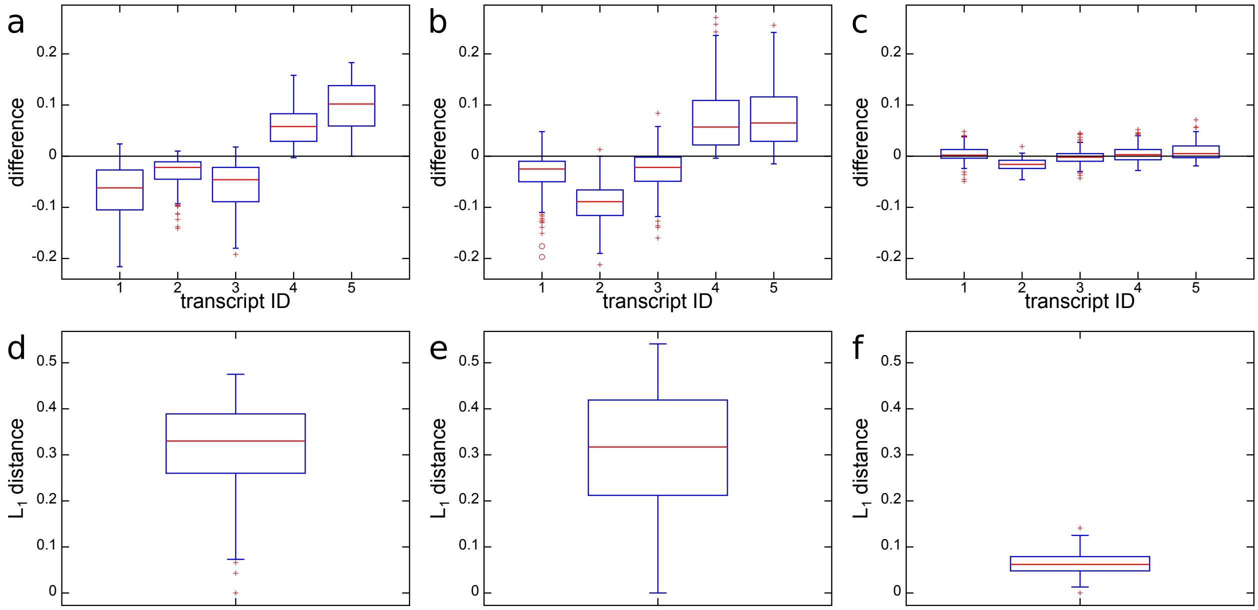

Figure 13: Difference between true and estimated abundances for **data with 5'+3' bias and correct annotations**. Figures (a), (b) and (c) show boxplots for the difference for each of the 5 transcripts in the **KLK5** gene for **Cufflinks 2.2.0** (a), **PennSeq** (b) and the **2p Mix^2 model with group tying** (c). Figures (d), (e) and (f) show boxplots of the L$_1$ distance for Cufflinks 2.2.0 (d), PennSeq (e) and the 2p Mix^2 model with group tying (f). The mean and the standard deviation of the values in Figures (c) and (d) are given in Table 11.

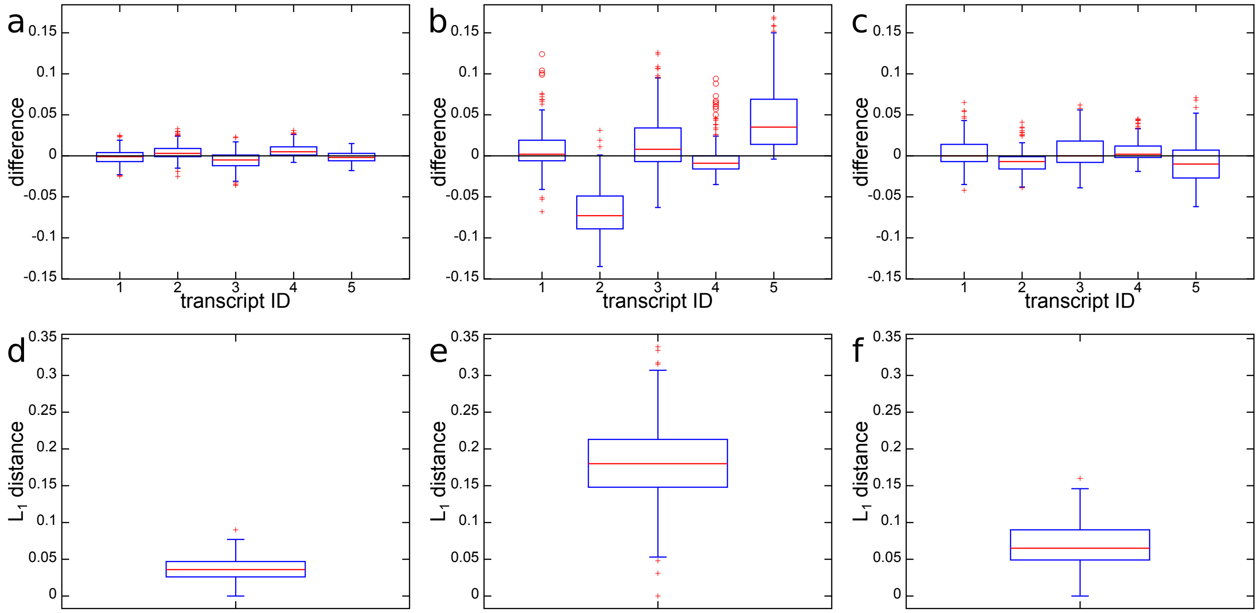

Figure 14: Difference between true and estimated abundances for **data with Cufflinks bias and correct annotations**. Figures (a), (b) and (c) show boxplots for the difference for each of the 5 transcripts in the **KLK5** gene for **Cufflinks 2.2.0** (a), **PennSeq** (b) and the **2p Mix^2 model with group tying** (c). Figures (d), (e) and (f) show boxplots of the L$_1$ distance for Cufflinks 2.2.0 (d), PennSeq (e) and the 2p Mix^2 model with group tying (f). The mean and the standard deviation of the values in Figures (c) and (d) are given in Table 12.

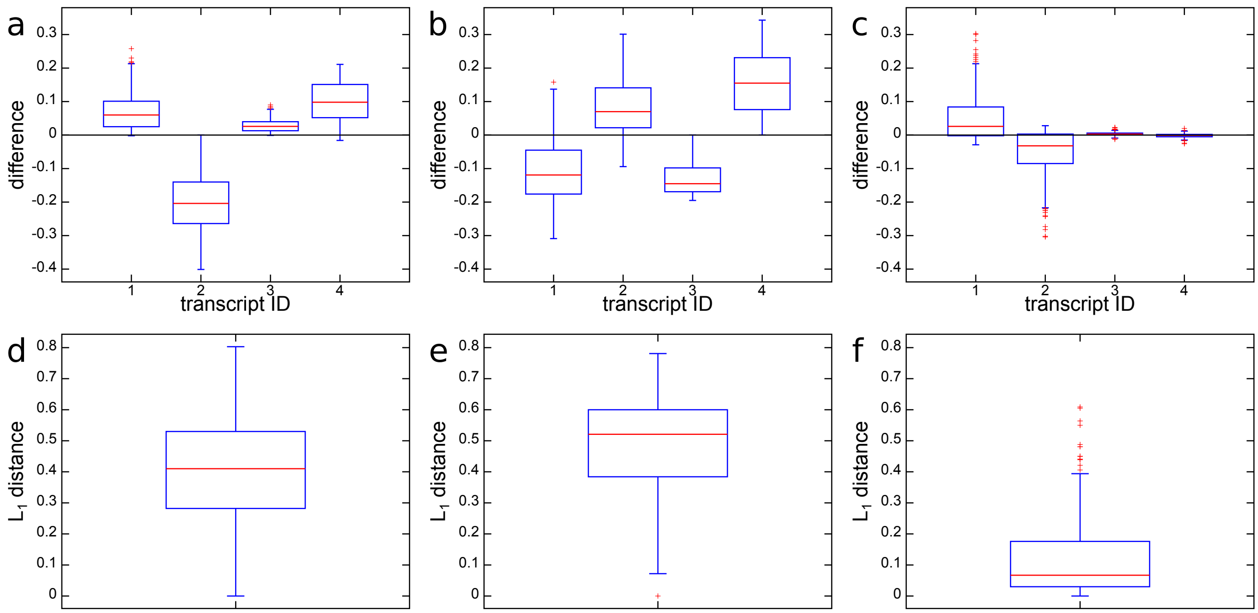

Figure 15: Difference between true and estimated abundances for **data with 5' bias and correct annotations**. Figures (a), (b) and (c) show boxplots for the difference for each of the 4 transcripts in the **LDHD** gene for **Cufflinks 2.2.0** (a), **PennSeq** (b) and the **2p Mix^2 model with group tying** (c). Figures (d), (e) and (f) show boxplots of the L_1 distance for Cufflinks 2.2.0 (d), PennSeq (e) and the 2p Mix^2 model with group tying (f). The mean and the standard deviation of the values in Figures (c) and (d) are given in Table 9.

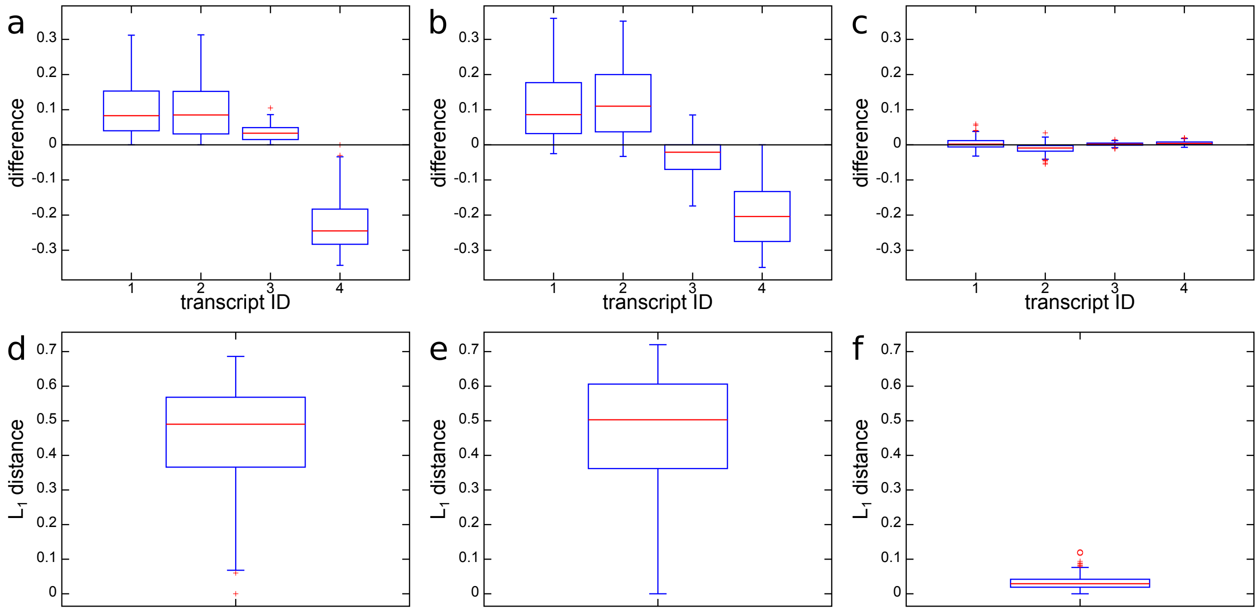

Figure 16: Difference between true and estimated abundances for **data with 3' bias and correct annotations**. Figures (a), (b) and (c) show boxplots for the difference for each of the 4 transcripts in the **LDHD** gene for **Cufflinks 2.2.0** (a), **PennSeq** (b) and the **2p Mix^2 model with group tying** (c). Figures (d), (e) and (f) show boxplots of the L_1 distance for Cufflinks 2.2.0 (d), PennSeq (e) and the 2p Mix^2 model with group tying (f). The mean and the standard deviation of the values in Figures (c) and (d) are given in Table 10.

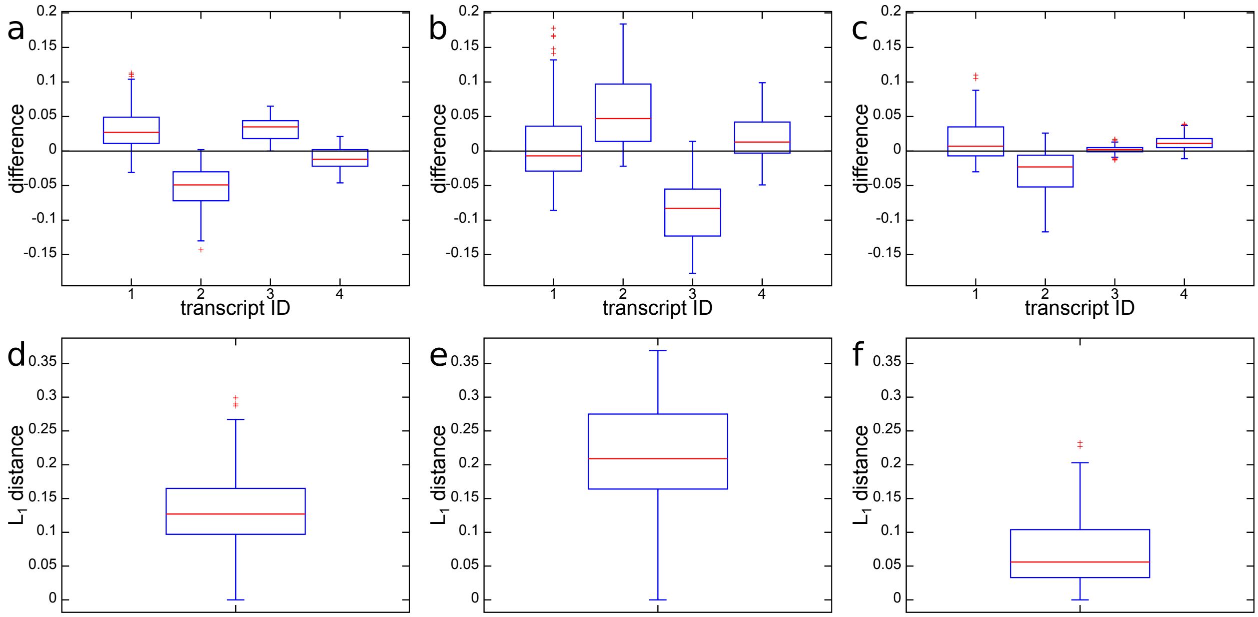

Figure 17: Difference between true and estimated abundances for **data with 5'+3' bias and correct annotations**. Figures (a), (b) and (c) show boxplots for the difference for each of the 4 transcripts in the **LDHD** gene for **Cufflinks 2.2.0** (a), **PennSeq** (b) and the **2p Mix^2 model with group tying** (c). Figures (d), (e) and (f) show boxplots of the L$_1$ distance for Cufflinks 2.2.0 (d), PennSeq (e) and the 2p Mix^2 model with group tying (f). The mean and the standard deviation of the values in Figures (c) and (d) are given in Table 11.

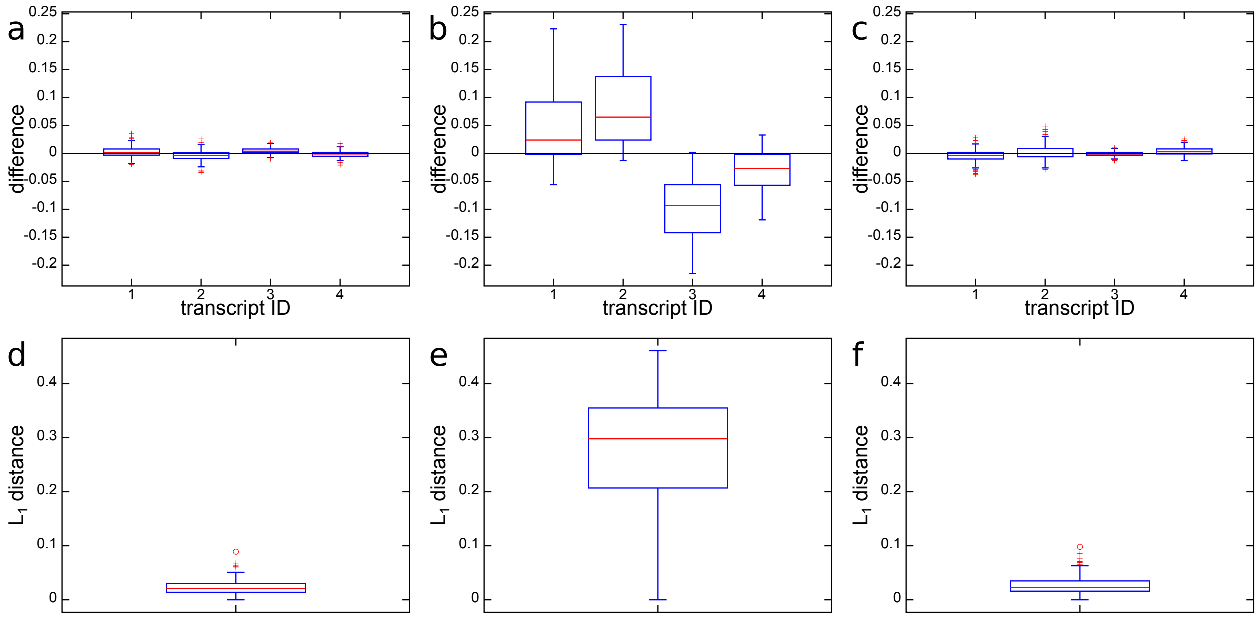

Figure 18: Difference between true and estimated abundances for **data with Cufflinks bias and correct annotations**. Figures (a), (b) and (c) show boxplots for the difference for each of the 4 transcripts in the **LDHD** gene for **Cufflinks 2.2.0** (a), **PennSeq** (b) and the **2p Mix^2 model with group tying** (c). Figures (d), (e) and (f) show boxplots of the L$_1$ distance for Cufflinks 2.2.0 (d), PennSeq (e) and the 2p Mix^2 model with group tying (f). The mean and the standard deviation of the values in Figures (c) and (d) are given in Table 12.

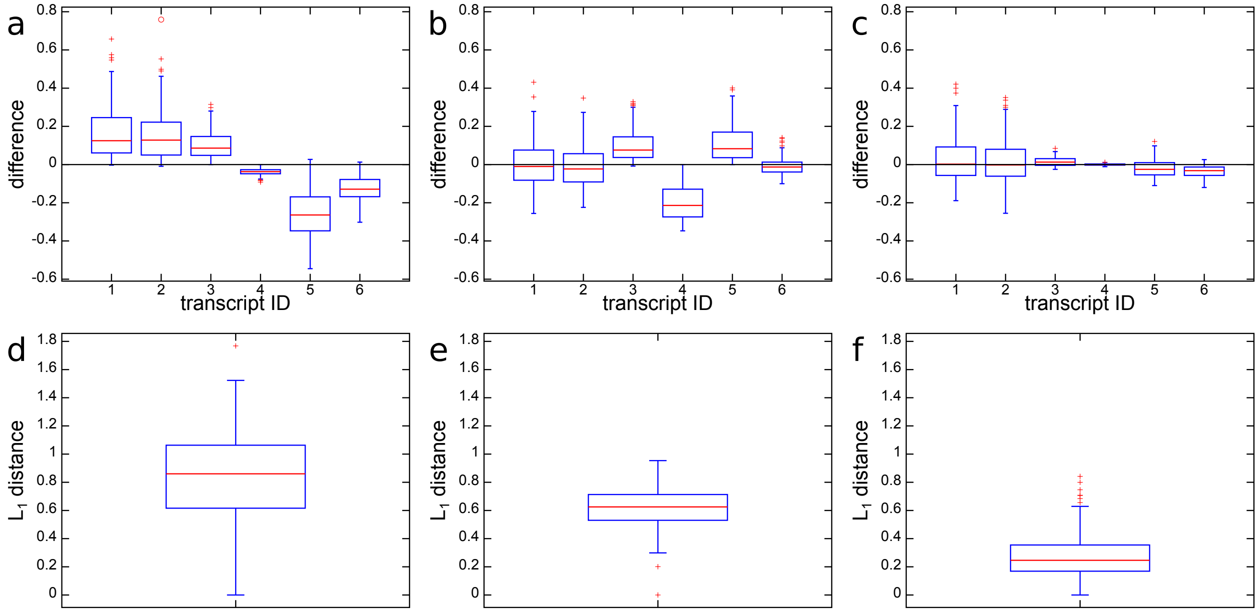

Figure 19: Difference between true and estimated abundances for **data with 5' bias and correct annotations**. Figures (a), (b) and (c) show boxplots for the difference for each of the 6 transcripts in the **LGALS17A** gene for **Cufflinks 2.2.0** (a), **PennSeq** (b) and the **2p Mix^2 model with group tying** (c). Figures (d), (e) and (f) show boxplots of the L$_1$ distance for Cufflinks 2.2.0 (d), PennSeq (e) and the 2p Mix^2 model with group tying (f). The mean and the standard deviation of the values in Figures (c) and (d) are given in Table 9.

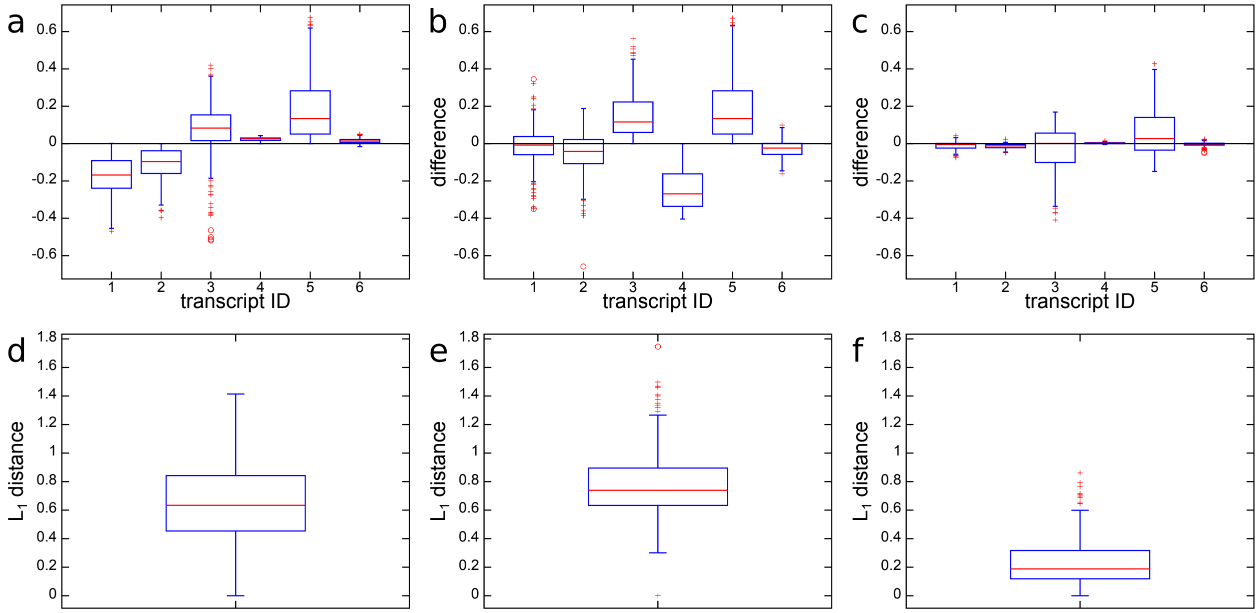

Figure 20: Difference between true and estimated abundances for **data with 3' bias and correct annotations**. Figures (a), (b) and (c) show boxplots for the difference for each of the 6 transcripts in the **LGALS17A** gene for **Cufflinks 2.2.0** (a), **PennSeq** (b) and the **2p Mix^2 model with group tying** (c). Figures (d), (e) and (f) show boxplots of the L$_1$ distance for Cufflinks 2.2.0 (d), PennSeq (e) and the 2p Mix^2 model with group tying (f). The mean and the standard deviation of the values in Figures (c) and (d) are given in Table 10.

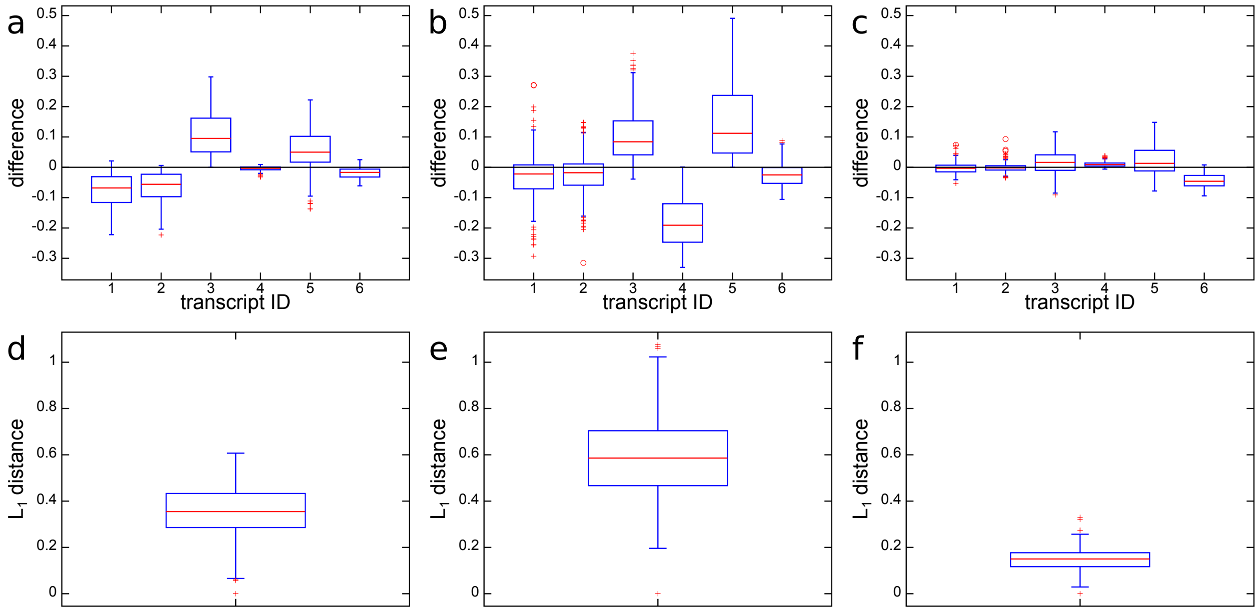

Figure 21: Difference between true and estimated abundances for **data with 5'+3' bias and correct annotations**. Figures (a), (b) and (c) show boxplots for the difference for each of the 6 transcripts in the **LGALS17A** gene for **Cufflinks 2.2.0** (a), **PennSeq** (b) and the **2p Mix^2 model with group tying** (c). Figures (d), (e) and (f) show boxplots of the L_1 distance for Cufflinks 2.2.0 (d), PennSeq (e) and the 2p Mix^2 model with group tying (f). The mean and the standard deviation of the values in Figures (c) and (d) are given in Table 11.

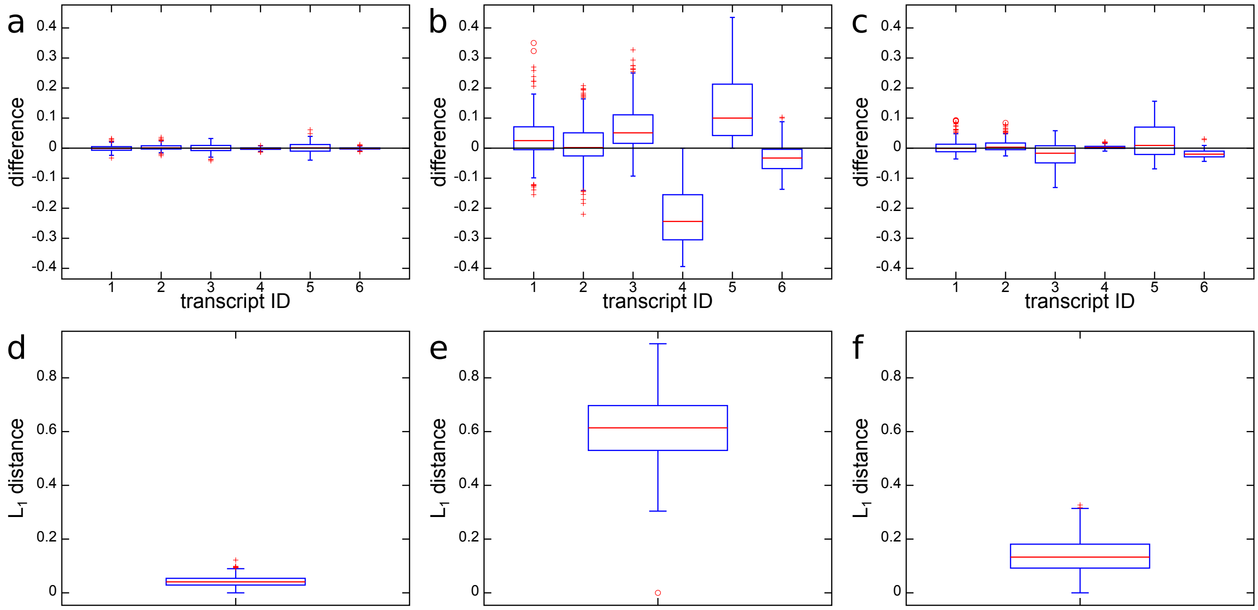

Figure 22: Difference between true and estimated abundances for **data with Cufflinks bias and correct annotations**. Figures (a), (b) and (c) show boxplots for the difference for each of the 6 transcripts in the **LGALS17A** gene for **Cufflinks 2.2.0** (a), **PennSeq** (b) and the **2p Mix^2 model with group tying** (c). Figures (d), (e) and (f) show boxplots of the L_1 distance for Cufflinks 2.2.0 (d), PennSeq (e) and the 2p Mix^2 model with group tying (f). The mean and the standard deviation of the values in Figures (c) and (d) are given in Table 12.

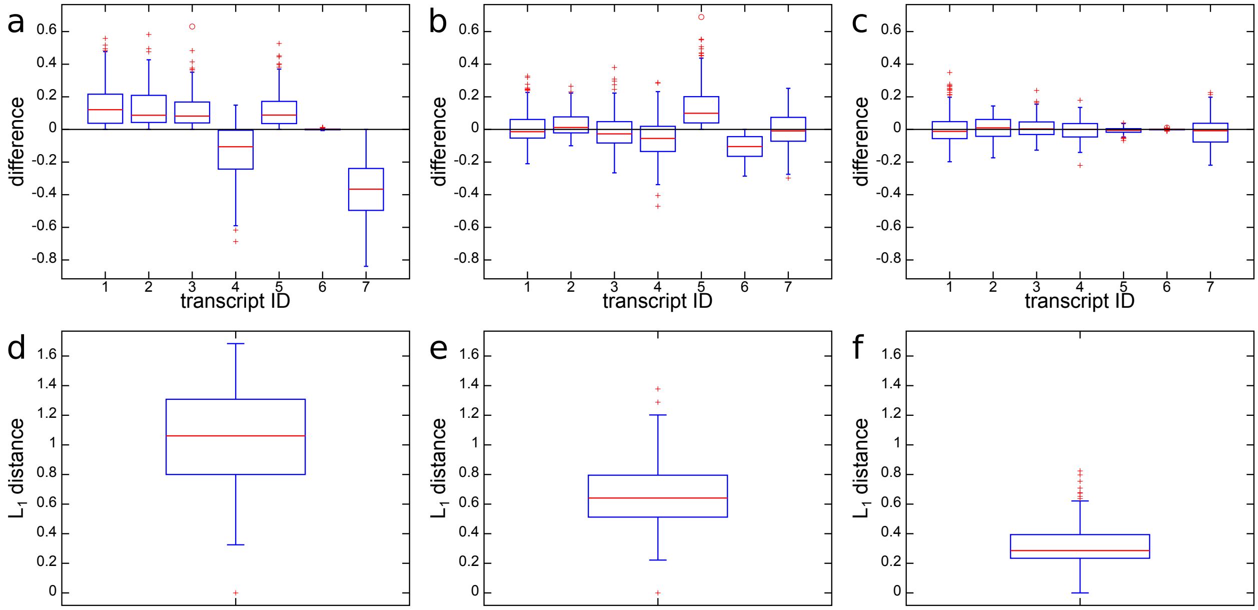

Figure 23: Difference between true and estimated abundances for **data with 5' bias and correct annotations**. Figures (a), (b) and (c) show boxplots for the difference for each of the 7 transcripts in the **DAPK3** gene for **Cufflinks 2.2.0** (a), **PennSeq** (b) and the **2p Mix^2 model with group tying** (c). Figures (d), (e) and (f) show boxplots of the L_1 distance for Cufflinks 2.2.0 (d), PennSeq (e) and the 2p Mix^2 model with group tying (f). The mean and the standard deviation of the values in Figures (c) and (d) are given in Table 9.

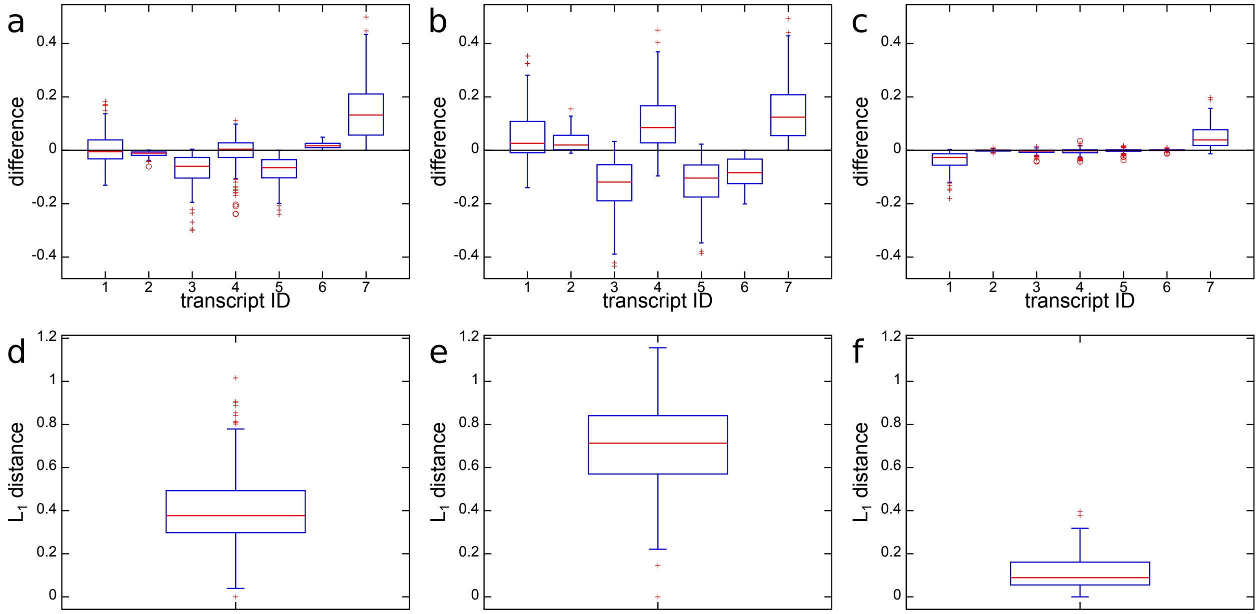

Figure 24: Difference between true and estimated abundances for **data with 3' bias and correct annotations**. Figures (a), (b) and (c) show boxplots for the difference for each of the 7 transcripts in the **DAPK3** gene for **Cufflinks 2.2.0** (a), **PennSeq** (b) and the **2p Mix^2 model with group tying** (c). Figures (d), (e) and (f) show boxplots of the L_1 distance for Cufflinks 2.2.0 (d), PennSeq (e) and the 2p Mix^2 model with group tying (f). The mean and the standard deviation of the values in Figures (c) and (d) are given in Table 10.

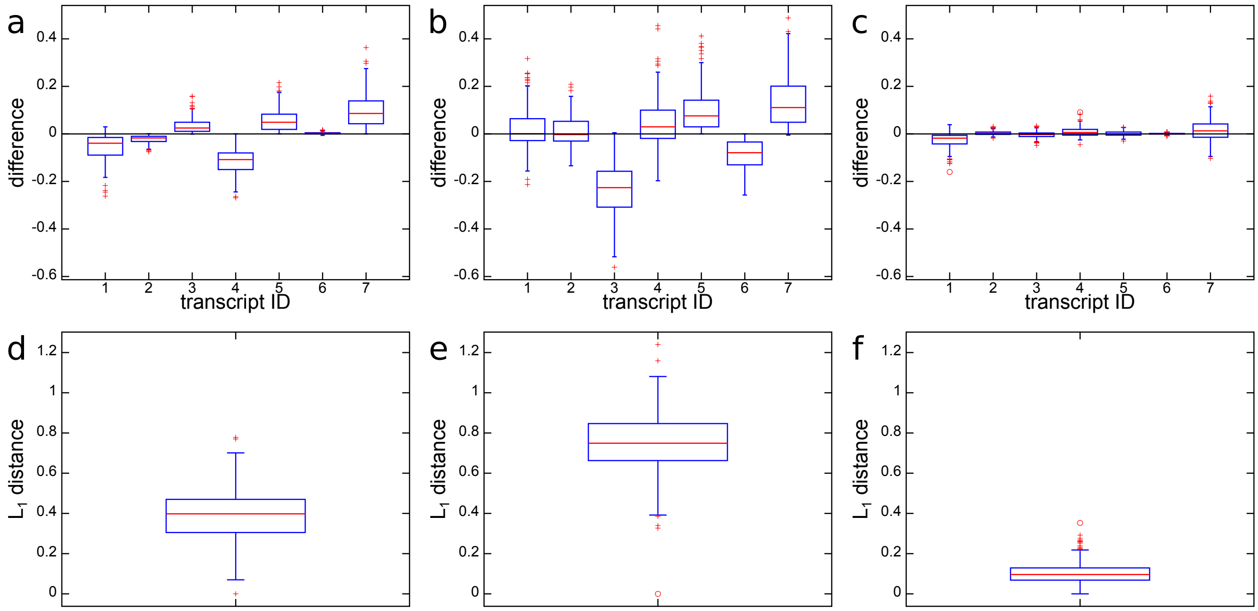

Figure 25: Difference between true and estimated abundances for **data with 5'+3' bias and correct annotations**. Figures (a), (b) and (c) show boxplots for the difference for each of the 7 transcripts in the **DAPK3** gene for **Cufflinks 2.2.0** (a), **PennSeq** (b) and the **2p Mix^2 model with group tying** (c). Figures (d), (e) and (f) show boxplots of the L_1 distance for Cufflinks 2.2.0 (d), PennSeq (e) and the 2p Mix^2 model with group tying (f). The mean and the standard deviation of the values in Figures (c) and (d) are given in Table 11.

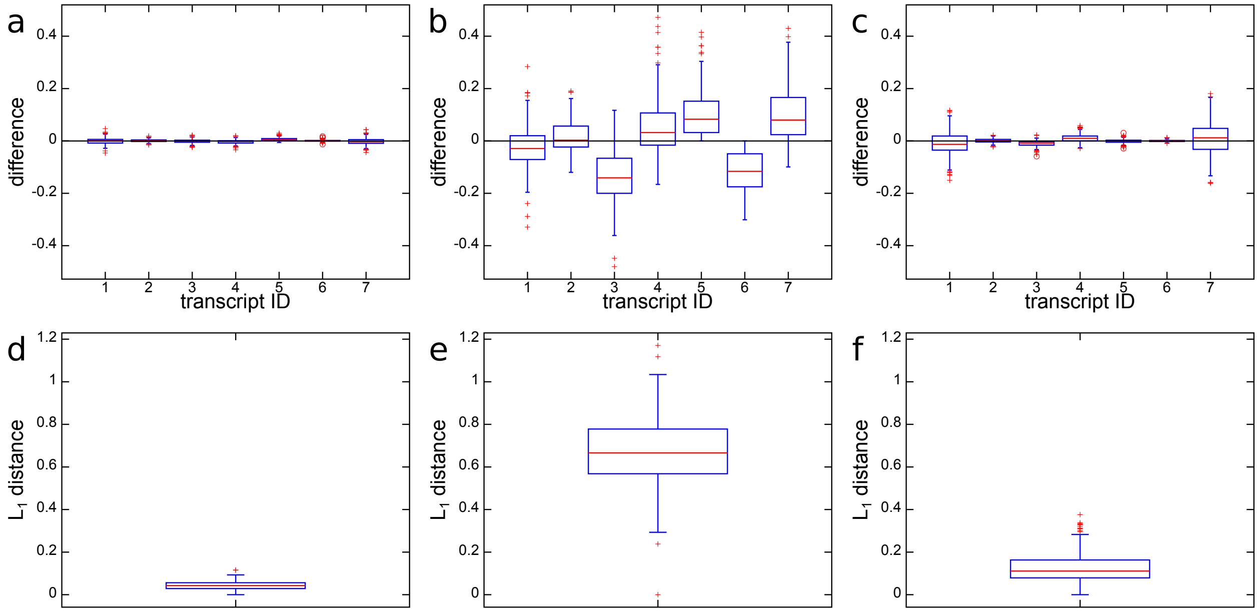

Figure 26: Difference between true and estimated abundances for **data with Cufflinks bias and correct annotations**. Figures (a), (b) and (c) show boxplots for the difference for each of the 7 transcripts in the **DAPK3** gene for **Cufflinks 2.2.0** (a), **PennSeq** (b) and the **2p Mix^2 model with group tying** (c). Figures (d), (e) and (f) show boxplots of the L_1 distance for Cufflinks 2.2.0 (d), PennSeq (e) and the 2p Mix^2 model with group tying (f). The mean and the standard deviation of the values in Figures (c) and (d) are given in Table 12.

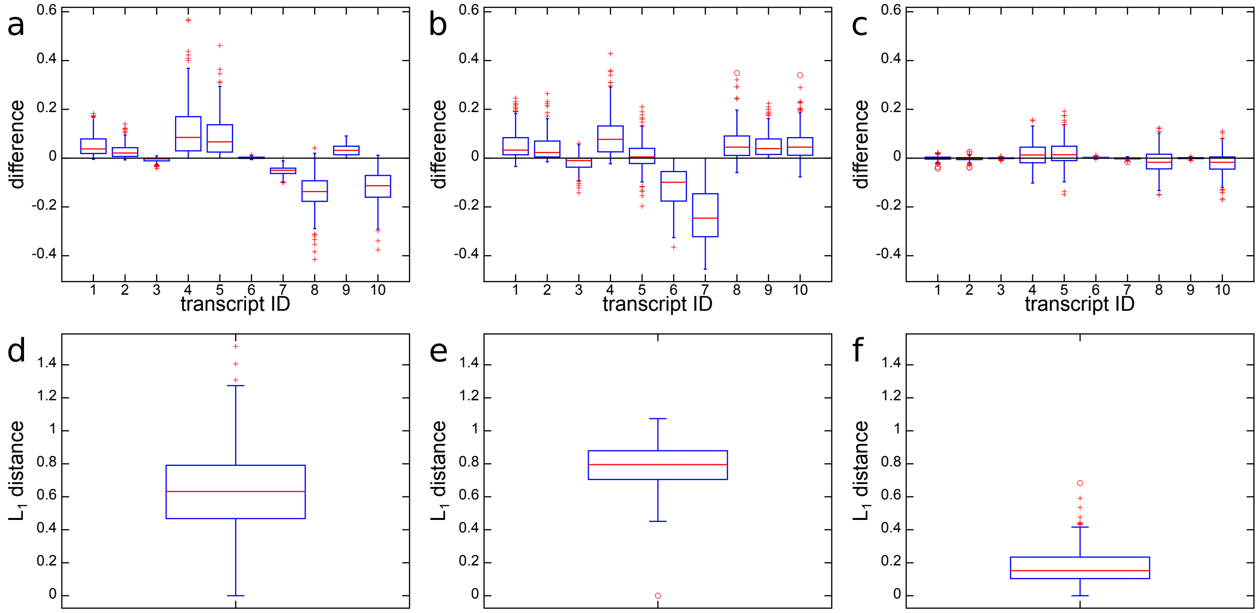

Figure 27: Difference between true and estimated abundances for **data with 5' bias and correct annotations**. Figures (a), (b) and (c) show boxplots for the difference for each of the 10 transcripts in the **HAUS5** gene for **Cufflinks 2.2.0** (a), **PennSeq** (b) and the **2p Mix^2 model with group tying** (c). Figures (d), (e) and (f) show boxplots of the L$_1$ distance for Cufflinks 2.2.0 (d), PennSeq (e) and the 2p Mix^2 model with group tying (f). The mean and the standard deviation of the values in Figures (c) and (d) are given in Table 9.

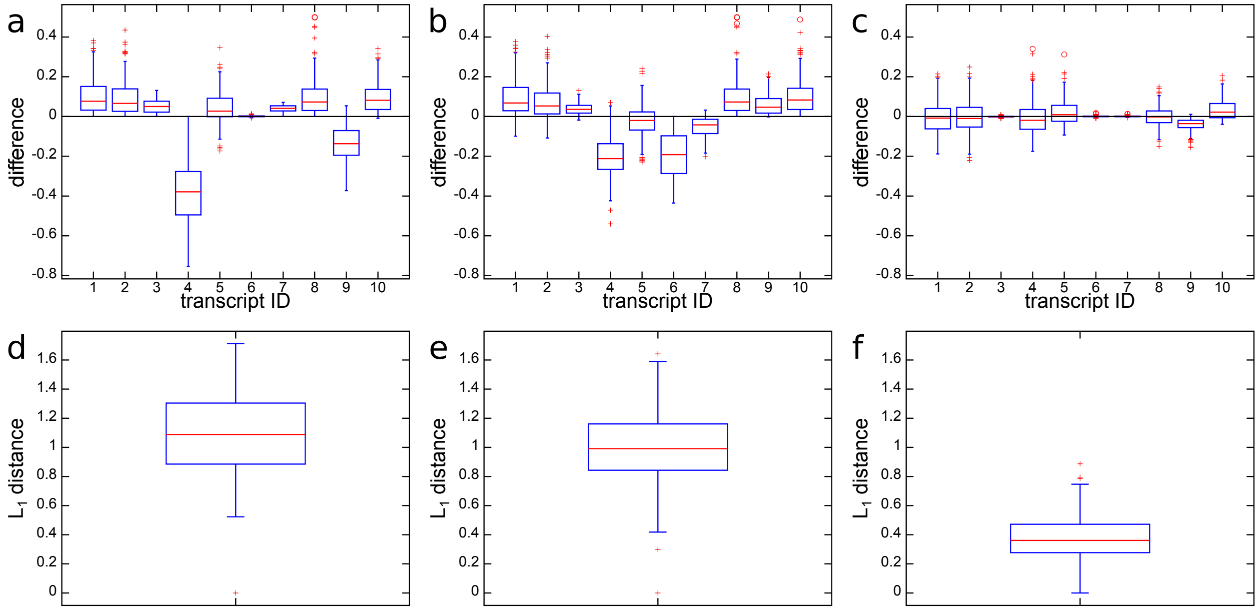

Figure 28: Difference between true and estimated abundances for **data with 3' bias and correct annotations**. Figures (a), (b) and (c) show boxplots for the difference for each of the 10 transcripts in the **HAUS5** gene for **Cufflinks 2.2.0** (a), **PennSeq** (b) and the **2p Mix^2 model with group tying** (c). Figures (d), (e) and (f) show boxplots of the L$_1$ distance for Cufflinks 2.2.0 (d), PennSeq (e) and the 2p Mix^2 model with group tying (f). The mean and the standard deviation of the values in Figures (c) and (d) are given in Table 10.

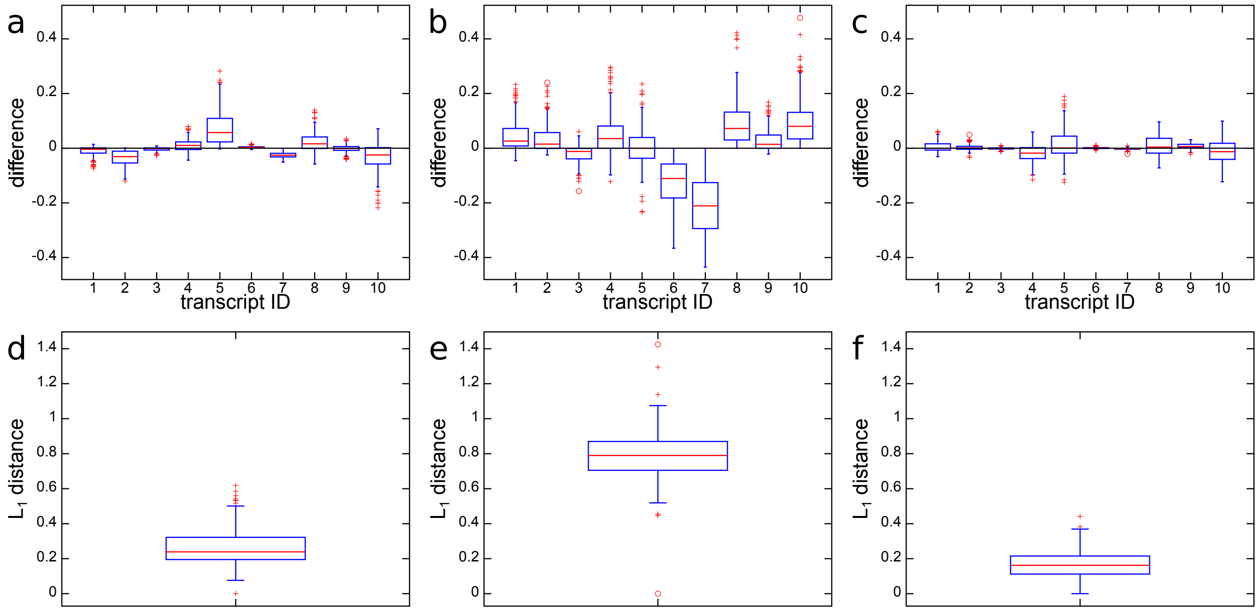

Figure 29: Difference between true and estimated abundances for **data with 5'+3' bias and correct annotations**. Figures (a), (b) and (c) show boxplots for the difference for each of the 10 transcripts in the **HAUS5** gene for **Cufflinks 2.2.0** (a), **PennSeq** (b) and the **2p Mix^2 model with group tying** (c). Figures (d), (e) and (f) show boxplots of the L_1 distance for Cufflinks 2.2.0 (d), PennSeq (e) and the 2p Mix^2 model with group tying (f). The mean and the standard deviation of the values in Figures (c) and (d) are given in Table 11.

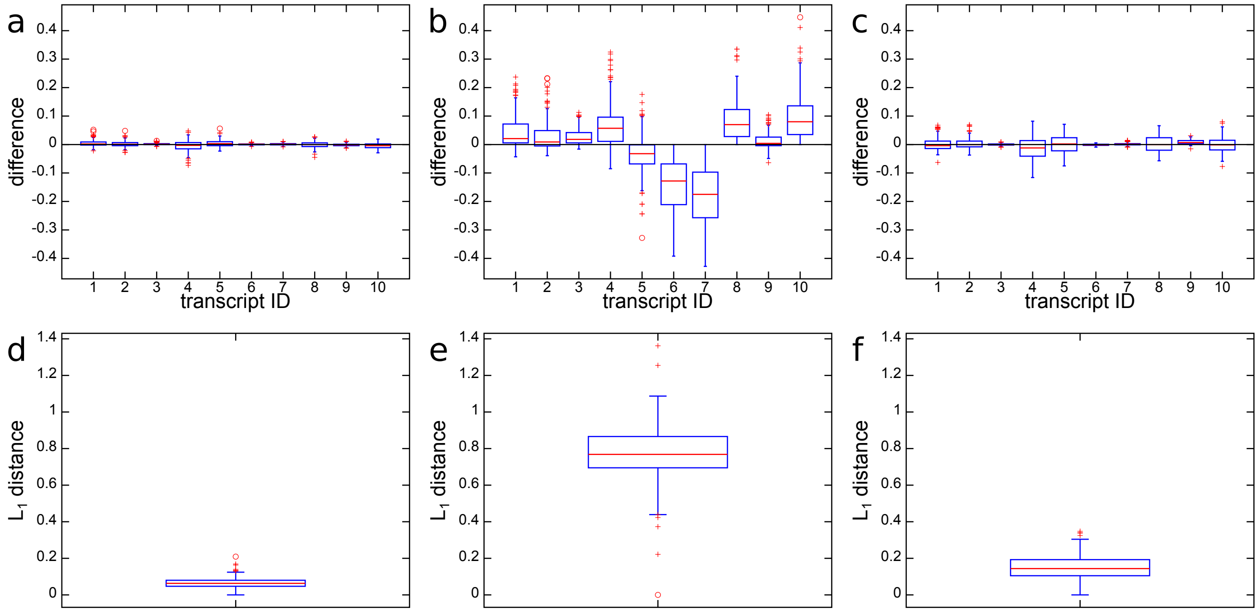

Figure 30: Difference between true and estimated abundances for **data with Cufflinks bias and correct annotations**. Figures (a), (b) and (c) show boxplots for the difference for each of the 10 transcripts in the **HAUS5** gene for **Cufflinks 2.2.0** (a), **PennSeq** (b) and the **2p Mix^2 model with group tying** (c). Figures (d), (e) and (f) show boxplots of the L_1 distance for Cufflinks 2.2.0 (d), PennSeq (e) and the 2p Mix^2 model with group tying (f). The mean and the standard deviation of the values in Figures (c) and (d) are given in Table 12.

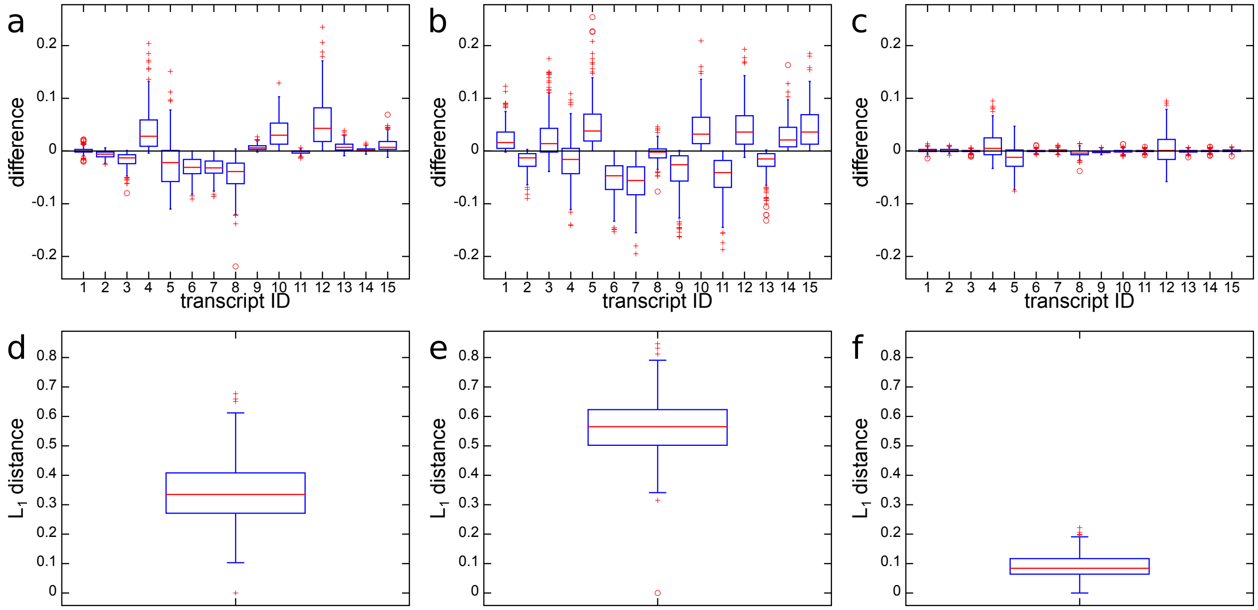

Figure 31: Difference between true and estimated abundances for **data with 5' bias and correct annotations**. Figures (a), (b) and (c) show boxplots for the difference for each of the 15 transcripts in the **USF2** gene for **Cufflinks 2.2.0** (a), **PennSeq** (b) and the **2p Mix^2 model with group tying** (c). Figures (d), (e) and (f) show boxplots of the L$_1$ distance for Cufflinks 2.2.0 (d), PennSeq (e) and the 2p Mix^2 model with group tying (f). The mean and the standard deviation of the values in Figures (c) and (d) are given in Table 9.

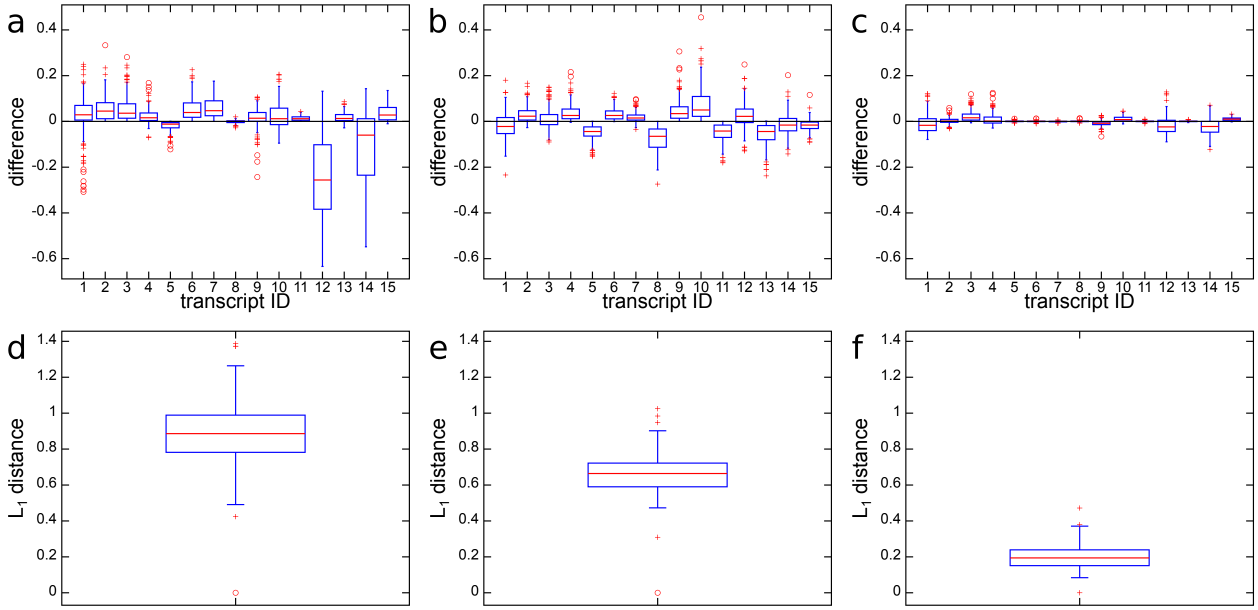

Figure 32: Difference between true and estimated abundances for **data with 3' bias and correct annotations**. Figures (a), (b) and (c) show boxplots for the difference for each of the 15 transcripts in the **USF2** gene for **Cufflinks 2.2.0** (a), **PennSeq** (b) and the **2p Mix^2 model with group tying** (c). Figures (d), (e) and (f) show boxplots of the L$_1$ distance for Cufflinks 2.2.0 (d), PennSeq (e) and the 2p Mix^2 model with group tying (f). The mean and the standard deviation of the values in Figures (c) and (d) are given in Table 10.

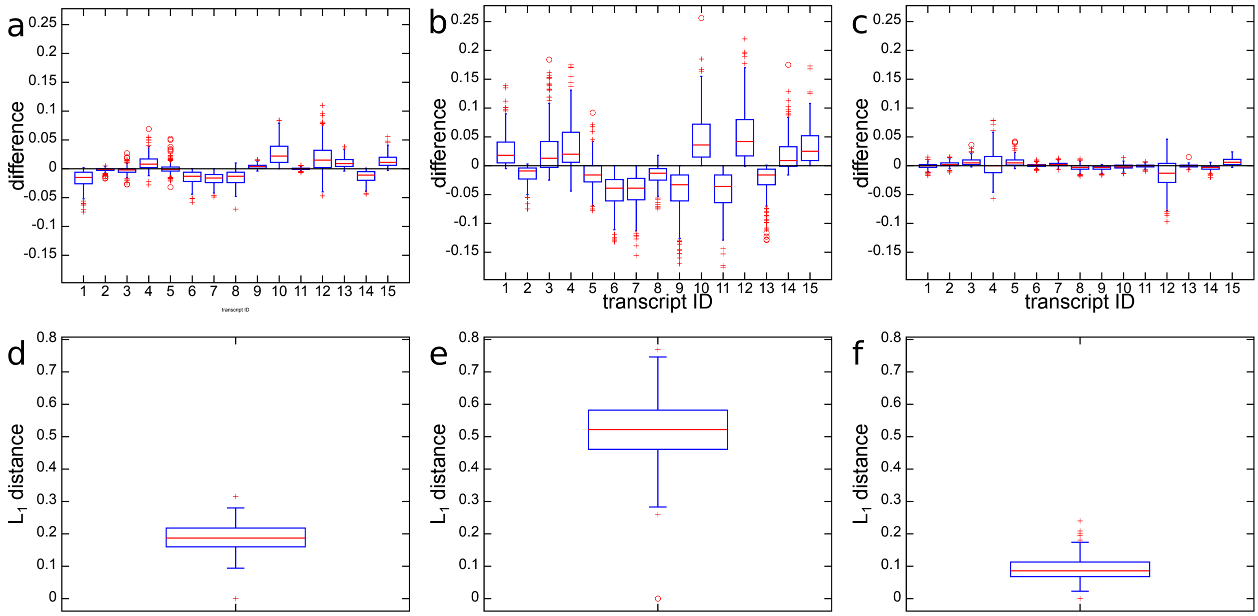

Figure 33: Difference between true and estimated abundances for **data with 5'+3' bias and correct annotations**. Figures (a), (b) and (c) show boxplots for the difference for each of the 15 transcripts in the **USF2** gene for **Cufflinks 2.2.0** (a), **PennSeq** (b) and the **2p Mix^2 model with group tying** (c). Figures (d), (e) and (f) show boxplots of the L_1 distance for Cufflinks 2.2.0 (d), PennSeq (e) and the 2p Mix^2 model with group tying (f). The mean and the standard deviation of the values in Figures (c) and (d) are given in Table 11.

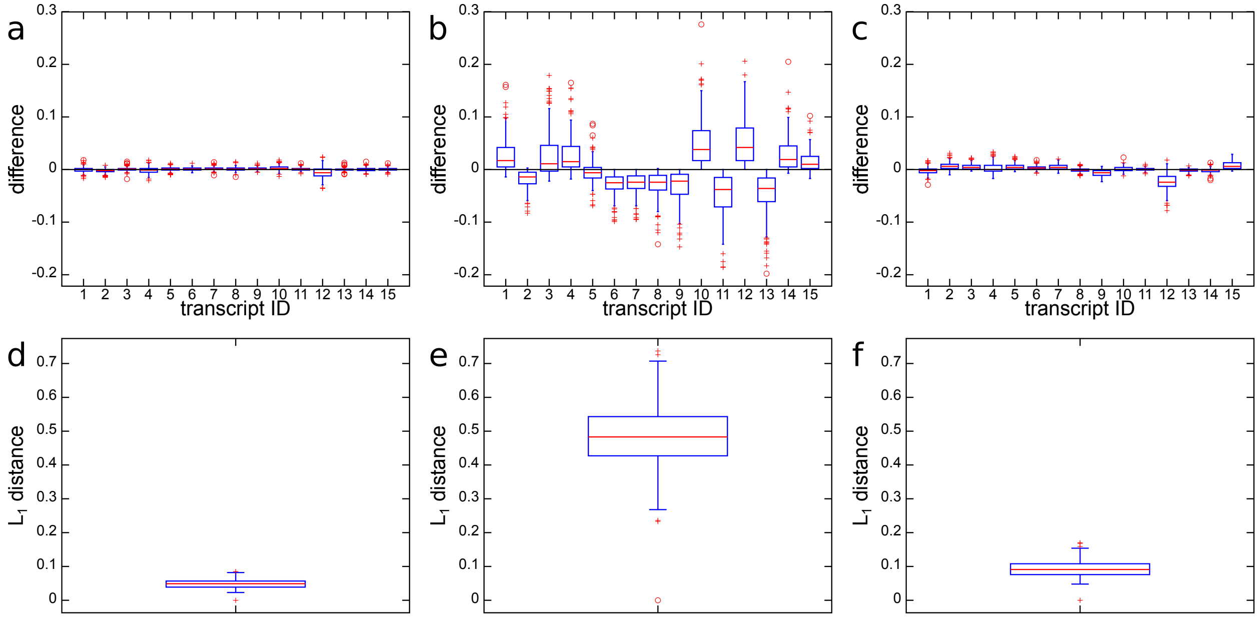

Figure 34: Difference between true and estimated abundances for **data with Cufflinks bias and correct annotations**. Figures (a), (b) and (c) show boxplots for the difference for each of the 15 transcripts in the **USF2** gene for **Cufflinks 2.2.0** (a), **PennSeq** (b) and the **2p Mix^2 model with group tying** (c). Figures (d), (e) and (f) show boxplots of the L_1 distance for Cufflinks 2.2.0 (d), PennSeq (e) and the 2p Mix^2 model with group tying (f). The mean and the standard deviation of the values in Figures (c) and (d) are given in Table 12.

5 Cufflinks, PennSeq and the Mix^2 model on incorrect annotations

This section contains the detailed results of the experiments on incorrect annotations comparing Cufflinks 2.2.0, PennSeq, the 2p Mix^2 model and the 4p Mix^2 model with group tying. The distribution of the difference between correct and incorrect transcript start and end annotations is visualized in Figure 5.

Figure 35 shows the average L_1 distance between true and estimated abundances for the 4 different biases, which, together with the standard deviations, are given in Table 15 to Table 18. Figure 7 in the main paper is a summary of Figure 35 in the supplement.

For each gene and each type of bias, boxplots of the difference as well as the L_1 distance between true and estimated abundances are given in Figure 36 to Figure 63. These boxplots summarize the results for Cufflinks 2.2.0, PennSeq, the 2p Mix^2 model and the 4p Mix^2 model with group tying for the 200 abundance sets that were sampled, according to the Dirichlet distribution, for each gene and bias. The 4p Mix^2 model with group tying consistently outperforms both Cufflinks 2.2.0 and PennSeq on all types of biases.

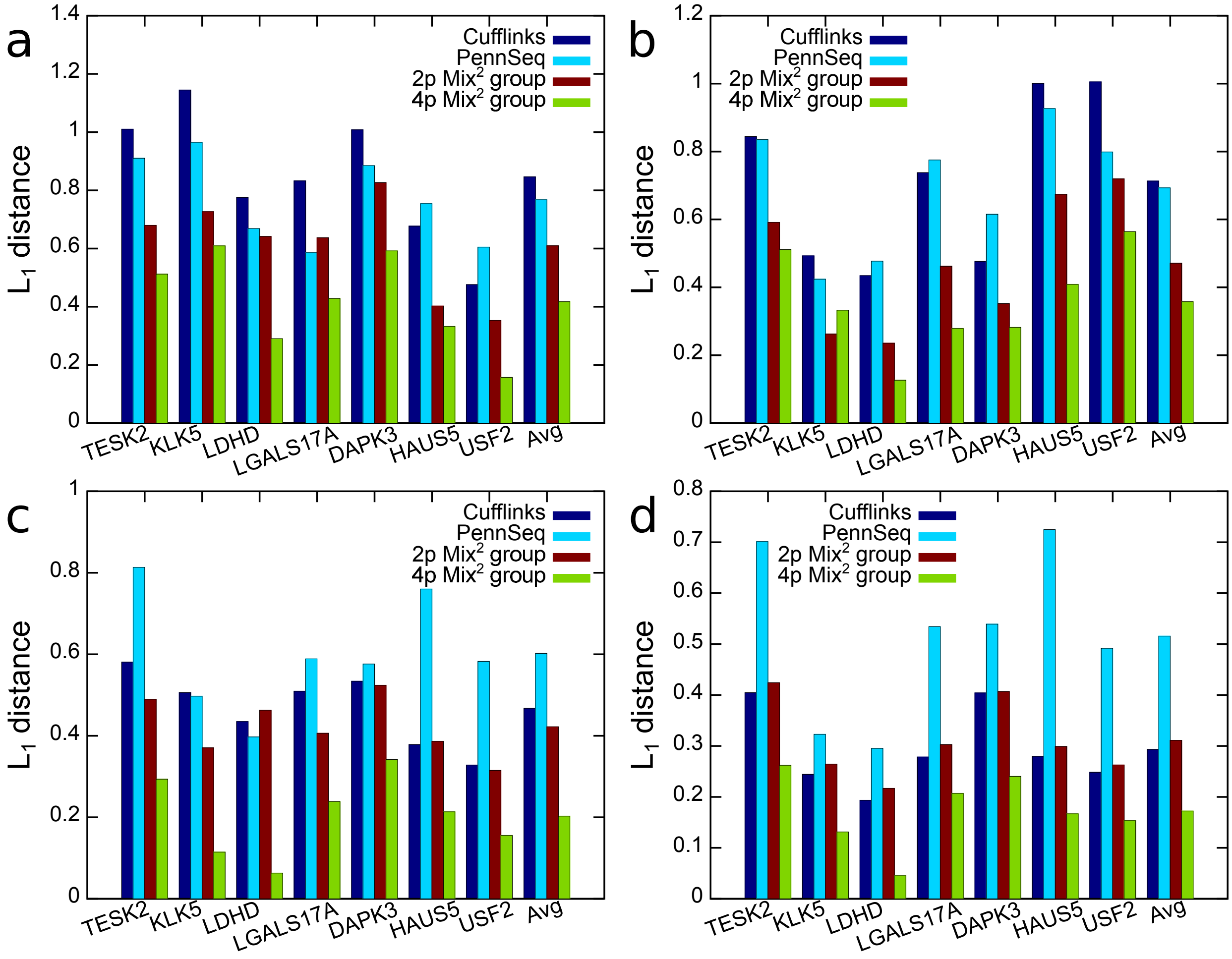

Figure 35: Average L_1 distance between true and estimated abundance with incorrect annotations and data with 5' bias (a), 3' bias (b), 5'+3' bias (c) and Cufflinks bias (d). The corresponding numbers, including the standard deviation of the L_1 distance can be found in Table 15 for data with 5' bias, in Table 16 for data with 3' bias, in Table 17 for data with 5'+3' bias and in Table 18 for data with Cufflinks bias.

Gene	Cufflinks		PennSeq		2p Mix^2 group		4p Mix^2 group	
	mean	std	mean	std	mean	std	mean	std
TESK2	1.009970	0.323490	0.910210	0.292560	0.679530	0.282310	0.512190	0.235070
KLK5	1.144410	0.332100	0.965200	0.364460	0.726750	0.324840	0.609270	0.218010
LDHD	0.775610	0.327430	0.668510	0.345600	0.641700	0.333240	0.290160	0.234870
LGALS17A	0.832620	0.300200	0.585790	0.177850	0.636980	0.276200	0.428400	0.194610
DAPK3	1.007820	0.335270	0.884850	0.303610	0.826520	0.300420	0.591940	0.241000
HAUS5	0.677280	0.240960	0.754210	0.149130	0.402450	0.222160	0.332380	0.163200
USF2	0.476020	0.151210	0.604750	0.138210	0.352520	0.161980	0.157250	0.078680

Table 15: Mean and standard deviation of L_1 distance between true and estimated abundance with incorrect annotations and data with 5' bias.

Gene	Cufflinks		PennSeq		2p Mix^2 group		4p Mix^2 group	
	mean	std	mean	std	mean	std	mean	std
TESK2	0.844130	0.236770	0.835010	0.265690	0.591040	0.271870	0.511360	0.208380
KLK5	0.492910	0.255930	0.424120	0.219020	0.262620	0.181190	0.332820	0.240780
LDHD	0.434290	0.171970	0.477160	0.183140	0.235610	0.191680	0.126530	0.059760
LGALS17A	0.737690	0.297150	0.775090	0.299970	0.461990	0.295480	0.279050	0.215190
DAPK3	0.476230	0.207560	0.615260	0.214930	0.352160	0.212060	0.282040	0.155900
HAUS5	1.000690	0.254830	0.926370	0.234630	0.674110	0.269890	0.408790	0.157150
USF2	1.005130	0.195490	0.798620	0.186170	0.719500	0.216560	0.563910	0.167140

Table 16: Mean and standard deviation of L_1 distance between true and estimated abundance with incorrect annotations and data with 3' bias.

Gene	Cufflinks		PennSeq		2p Mix^2 group		4p Mix^2 group	
	mean	std	mean	std	mean	std	mean	std
TESK2	0.580780	0.247050	0.813340	0.216990	0.489560	0.217720	0.293540	0.177070
KLK5	0.506000	0.206020	0.497180	0.198360	0.370480	0.200530	0.114520	0.078990
LDHD	0.434880	0.211350	0.397300	0.210790	0.462710	0.254420	0.062990	0.024880
LGALS17A	0.509180	0.203940	0.588590	0.221980	0.406140	0.205230	0.238560	0.088340
DAPK3	0.533860	0.250560	0.576060	0.219010	0.523700	0.238360	0.341770	0.179960
HAUS5	0.378540	0.146530	0.760010	0.158950	0.386430	0.174240	0.213430	0.109880
USF2	0.328170	0.110190	0.582600	0.130810	0.314870	0.124140	0.155350	0.054990

Table 17: Mean and standard deviation of L_1 distance between true and estimated abundance with incorrect annotations and data with 5'+3' bias.

Gene	Cufflinks		PennSeq		2p Mix^2 group		4p Mix^2 group	
	mean	std	mean	std	mean	std	mean	std
TESK2	0.404780	0.192760	0.701170	0.197900	0.424160	0.216080	0.262130	0.135540
KLK5	0.244220	0.136130	0.322890	0.161180	0.264300	0.142650	0.131120	0.088180
LDHD	0.193240	0.102810	0.295260	0.125790	0.216720	0.125920	0.045170	0.023720
LGALS17A	0.278250	0.135250	0.534230	0.171130	0.302810	0.155180	0.206990	0.108000
DAPK3	0.404250	0.213100	0.539280	0.222480	0.406870	0.238770	0.240290	0.149820
HAUS5	0.279750	0.130100	0.724830	0.165860	0.298990	0.131270	0.166750	0.098740
USF2	0.248320	0.085720	0.491700	0.119580	0.262620	0.106790	0.153360	0.053260

Table 18: Mean and standard deviation of L_1 distance between true and estimated abundance with incorrect annotations and data with Cufflinks bias.

Gene	Cufflinks		PennSeq		2p Mix^2 group		4p Mix^2 group	
	mean	std	mean	std	mean	std	mean	std
TESK2	0.70992	0.25002	0.81493	0.24328	0.54607	0.24700	0.39481	0.18901
KLK5	0.59688	0.23255	0.55235	0.23575	0.40604	0.21230	0.29693	0.15649
LDHD	0.45950	0.20339	0.45956	0.21633	0.38919	0.22631	0.13121	0.08581
LGALS17A	0.58943	0.23414	0.62092	0.21773	0.45198	0.23302	0.28825	0.15154
DAPK3	0.60554	0.25162	0.65386	0.24001	0.52731	0.24740	0.36401	0.18167
HAUS5	0.58407	0.19310	0.79135	0.17714	0.44049	0.19939	0.28034	0.13224
USF2	0.51441	0.13565	0.61942	0.14369	0.41238	0.15237	0.25747	0.08852

Table 19: Mean and standard deviation of L_1 distance between true and estimated abundance with incorrect annotations averaged over biases. The means are visualized in Figure 7(a) in the main paper.

bias	Cufflinks		PennSeq		2p Mix^2 group		4p Mix^2 group	
	mean	std	mean	std	mean	std	mean	std
5' bias	0.84625	0.28724	0.76765	0.25306	0.60949	0.27159	0.41737	0.19506
3' bias	0.71301	0.23139	0.69309	0.22908	0.47100	0.23410	0.35779	0.17204
5'+3' bias	0.46734	0.19652	0.60215	0.19384	0.42198	0.20209	0.20288	0.10202
Cufflinks bias	0.29326	0.14227	0.51562	0.16627	0.31092	0.15952	0.17226	0.09389

Table 20: Mean and standard deviation of L_1 distance between true and estimated abundance with incorrect annotations averaged over genes. The means are visualized in Figure 7(b) in the main paper.

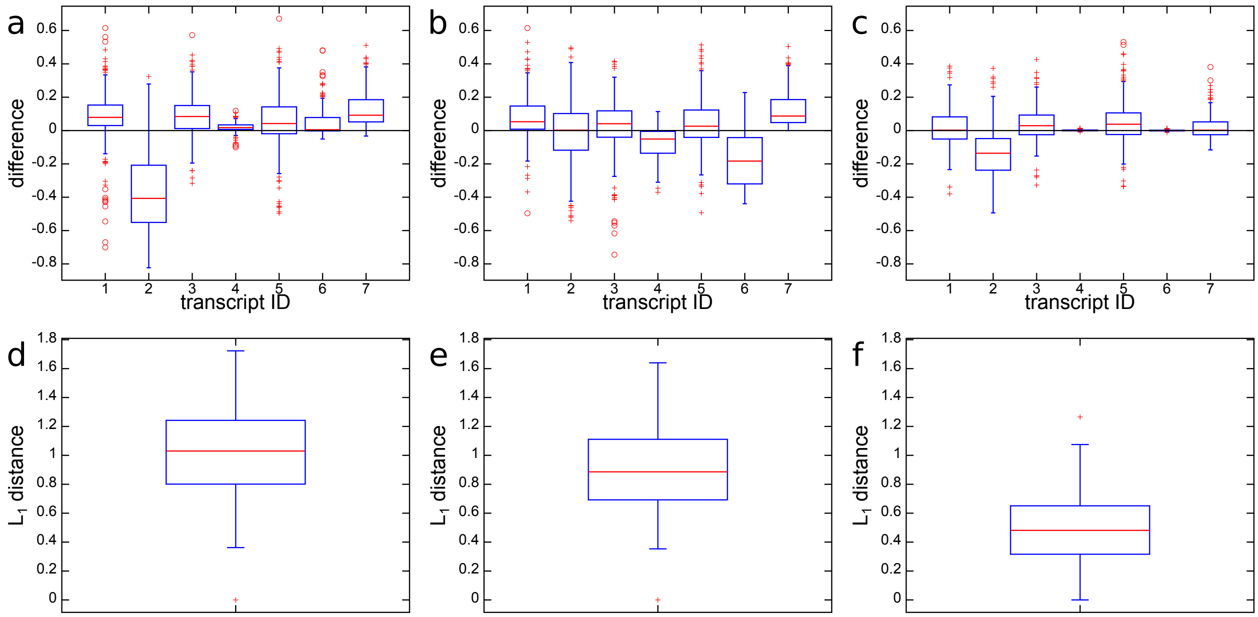

Figure 36: Difference between true and estimated abundances for **data with 5' bias and incorrect annotations**. Figures (a), (b) and (c) show boxplots for the difference for each of the 7 transcripts in the **TESK2** gene for **Cufflinks 2.2.0** (a), **PennSeq** (b) and the **4p Mix^2 model with group tying** (c). Figures (d), (e) and (f) show boxplots of the L_1 distance for Cufflinks 2.2.0 (d), PennSeq (e) and the 4p Mix^2 model with group tying (f). The mean and the standard deviation of the values in Figures (c) and (d) are given in Table 15.

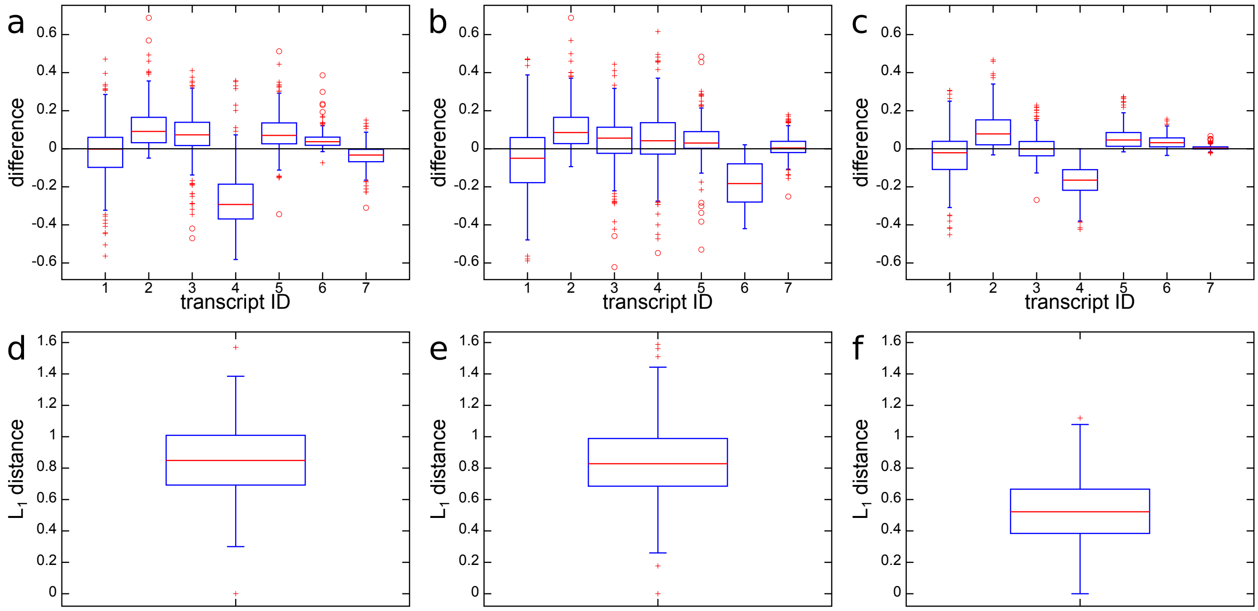

Figure 37: Difference between true and estimated abundances for **data with 3' bias and incorrect annotations**. Figures (a), (b) and (c) show boxplots for the difference for each of the 7 transcripts in the **TESK2** gene for **Cufflinks 2.2.0** (a), **PennSeq** (b) and the **4p Mix^2 model with group tying** (c). Figures (d), (e) and (f) show boxplots of the L_1 distance for Cufflinks 2.2.0 (d), PennSeq (e) and the 4p Mix^2 model with group tying (f). The mean and the standard deviation of the values in Figures (c) and (d) are given in Table 16.

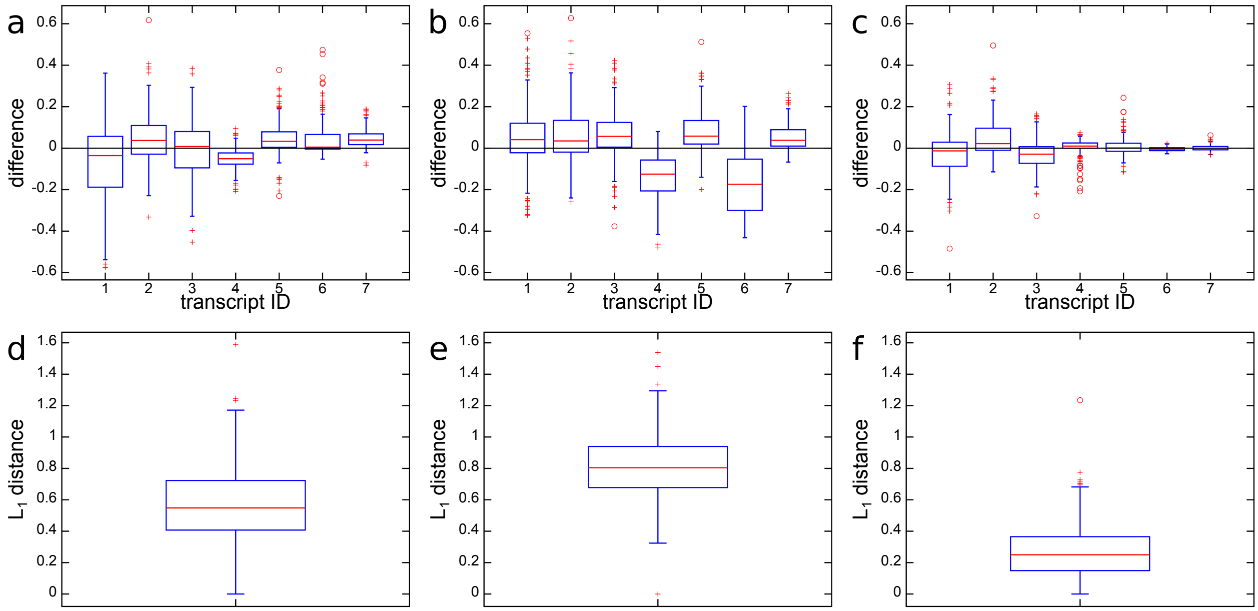

Figure 38: Difference between true and estimated abundances for **data with 5'+3' bias and incorrect annotations**. Figures (a), (b) and (c) show boxplots for the difference for each of the 7 transcripts in the **TESK2** gene for **Cufflinks 2.2.0** (a), **PennSeq** (b) and the **4p Mix^2 model with group tying** (c). Figures (d), (e) and (f) show boxplots of the L_1 distance for Cufflinks 2.2.0 (d), PennSeq (e) and the 4p Mix^2 model with group tying (f). The mean and the standard deviation of the values in Figures (c) and (d) are given in Table 17.

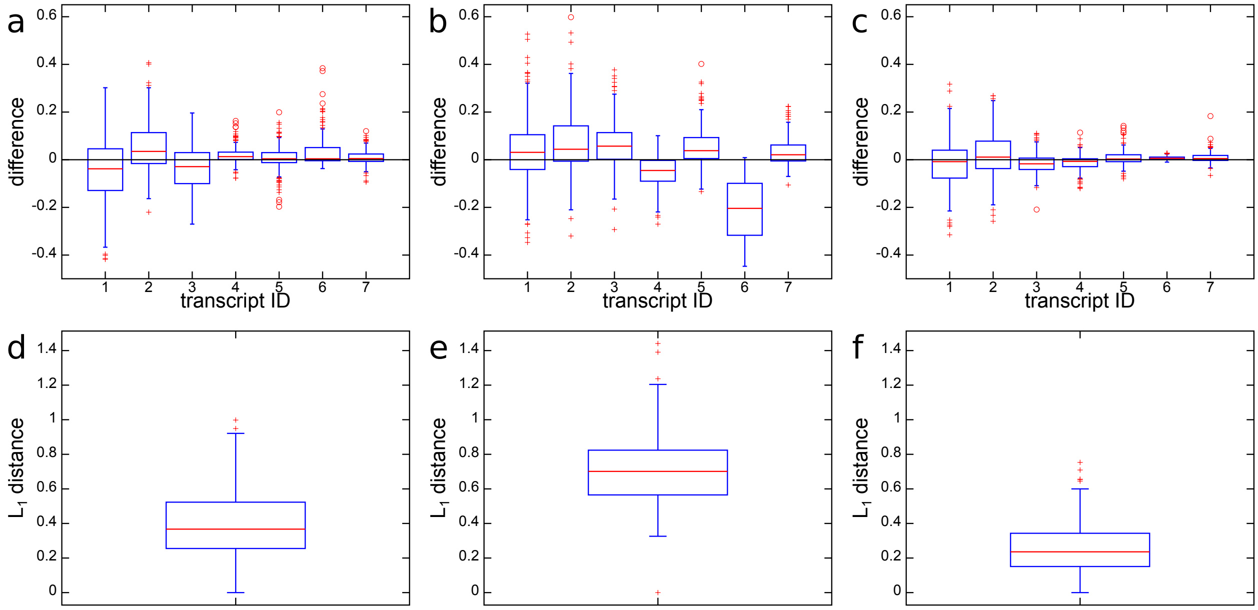

Figure 39: Difference between true and estimated abundances for **data with Cufflinks bias and incorrect annotations**. Figures (a), (b) and (c) show boxplots for the difference for each of the 7 transcripts in the **TESK2** gene for **Cufflinks 2.2.0** (a), **PennSeq** (b) and the **4p Mix^2 model with group tying** (c). Figures (d), (e) and (f) show boxplots of the L_1 distance for Cufflinks 2.2.0 (d), PennSeq (e) and the 4p Mix^2 model with group tying (f). The mean and the standard deviation of the values in Figures (c) and (d) are given in Table 18.

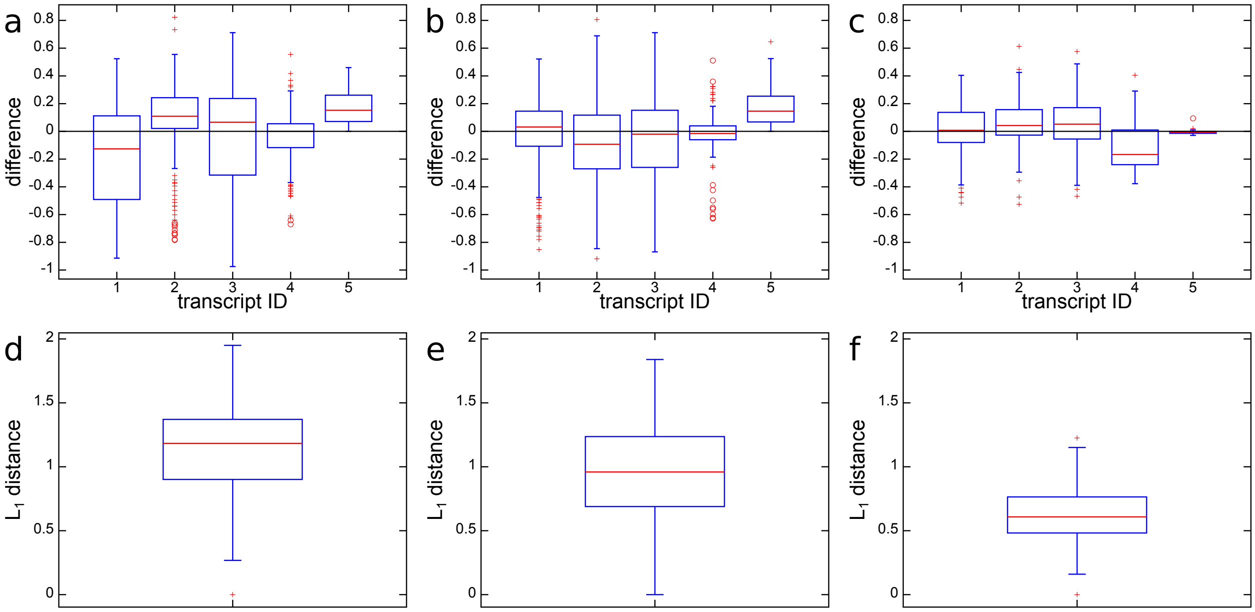

Figure 40: Difference between true and estimated abundances for **data with 5' bias and incorrect annotations**. Figures (a), (b) and (c) show boxplots for the difference for each of the 5 transcripts in the **KLK5** gene for **Cufflinks 2.2.0** (a), **PennSeq** (b) and the **4p Mix^2 model with group tying** (c). Figures (d), (e) and (f) show boxplots of the L$_1$ distance for Cufflinks 2.2.0 (d), PennSeq (e) and the 4p Mix^2 model with group tying (f). The mean and the standard deviation of the values in Figures (c) and (d) are given in Table 15.

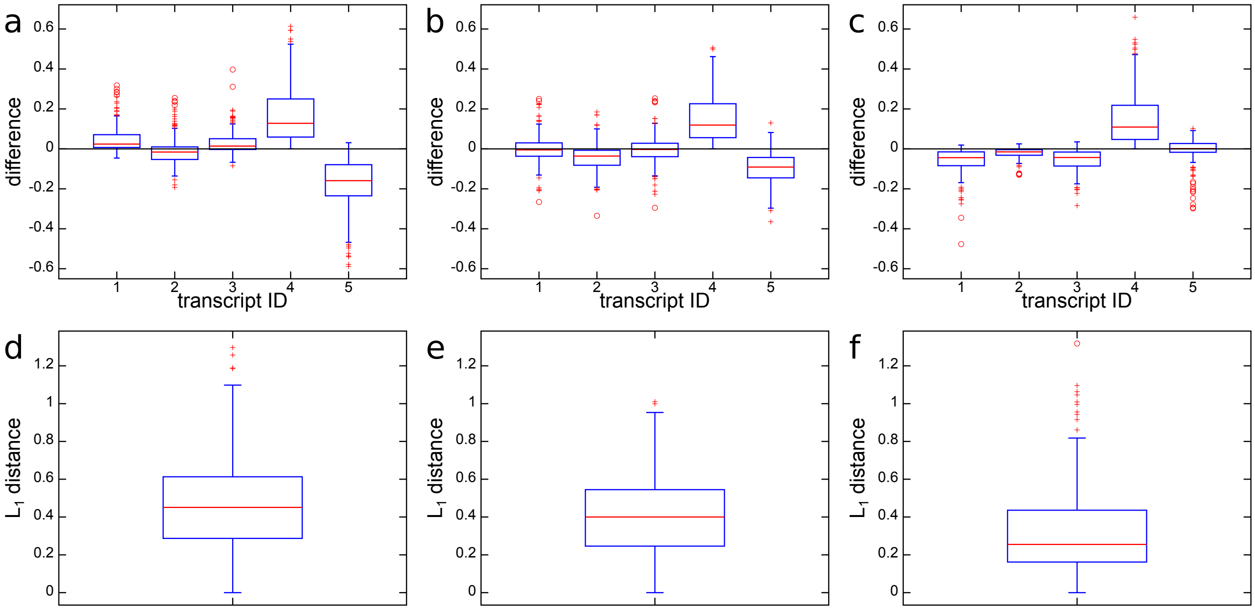

Figure 41: Difference between true and estimated abundances for **data with 3' bias and incorrect annotations**. Figures (a), (b) and (c) show boxplots for the difference for each of the 5 transcripts in the **KLK5** gene for **Cufflinks 2.2.0** (a), **PennSeq** (b) and the **4p Mix^2 model with group tying** (c). Figures (d), (e) and (f) show boxplots of the L$_1$ distance for Cufflinks 2.2.0 (d), PennSeq (e) and the 4p Mix^2 model with group tying (f). The mean and the standard deviation of the values in Figures (c) and (d) are given in Table 16.

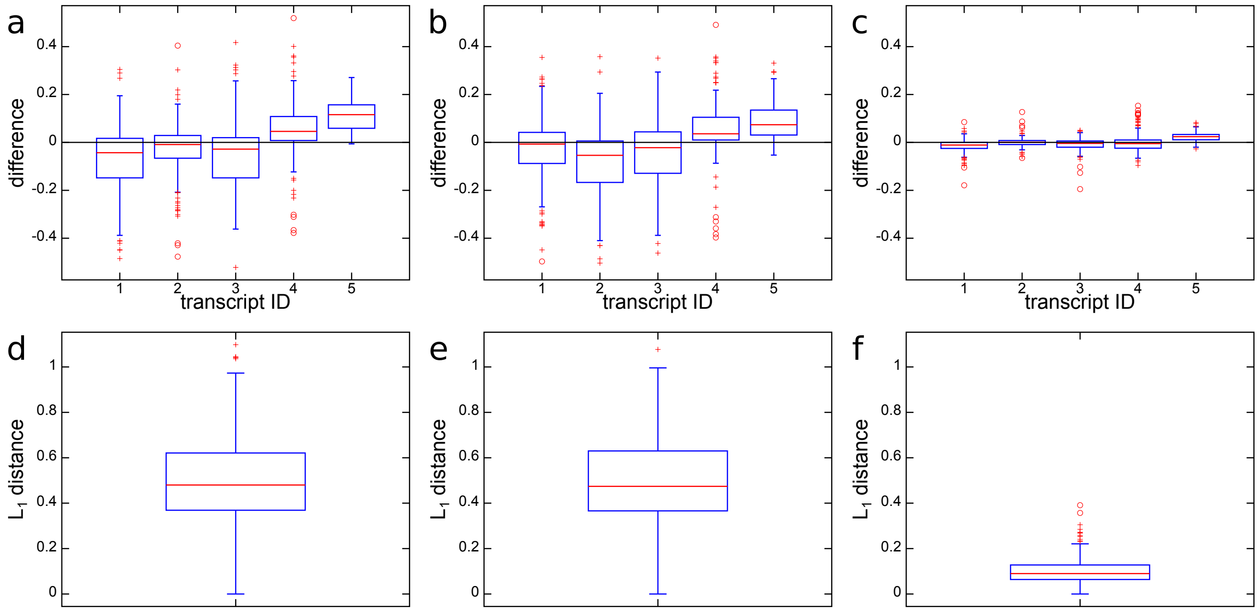

Figure 42: Difference between true and estimated abundances for **data with 5'+3' bias and incorrect annotations**. Figures (a), (b) and (c) show boxplots for the difference for each of the 5 transcripts in the **KLK5** gene for **Cufflinks 2.2.0** (a), **PennSeq** (b) and the **4p Mix^2 model with group tying** (c). Figures (d), (e) and (f) show boxplots of the L_1 distance for Cufflinks 2.2.0 (d), PennSeq (e) and the 4p Mix^2 model with group tying (f). The mean and the standard deviation of the values in Figures (c) and (d) are given in Table 17.

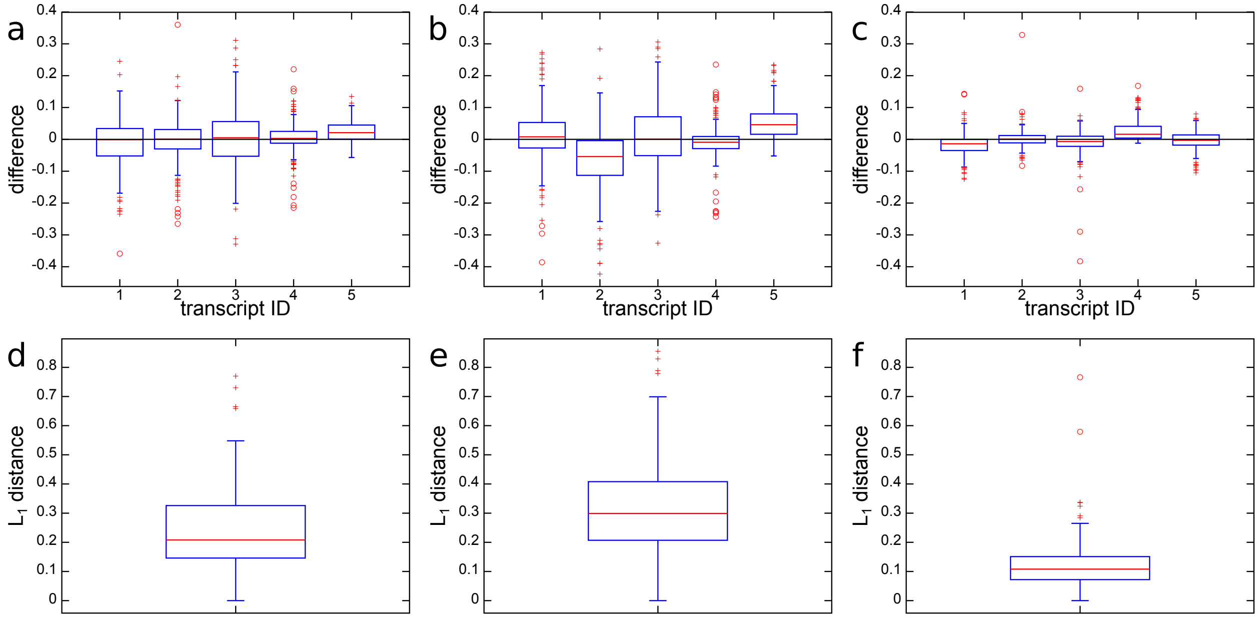

Figure 43: Difference between true and estimated abundances for **data with Cufflinks bias and incorrect annotations**. Figures (a), (b) and (c) show boxplots for the difference for each of the 5 transcripts in the **KLK5** gene for **Cufflinks 2.2.0** (a), **PennSeq** (b) and the **4p Mix^2 model with group tying** (c). Figures (d), (e) and (f) show boxplots of the L_1 distance for Cufflinks 2.2.0 (d), PennSeq (e) and the 4p Mix^2 model with group tying (f). The mean and the standard deviation of the values in Figures (c) and (d) are given in Table 18.

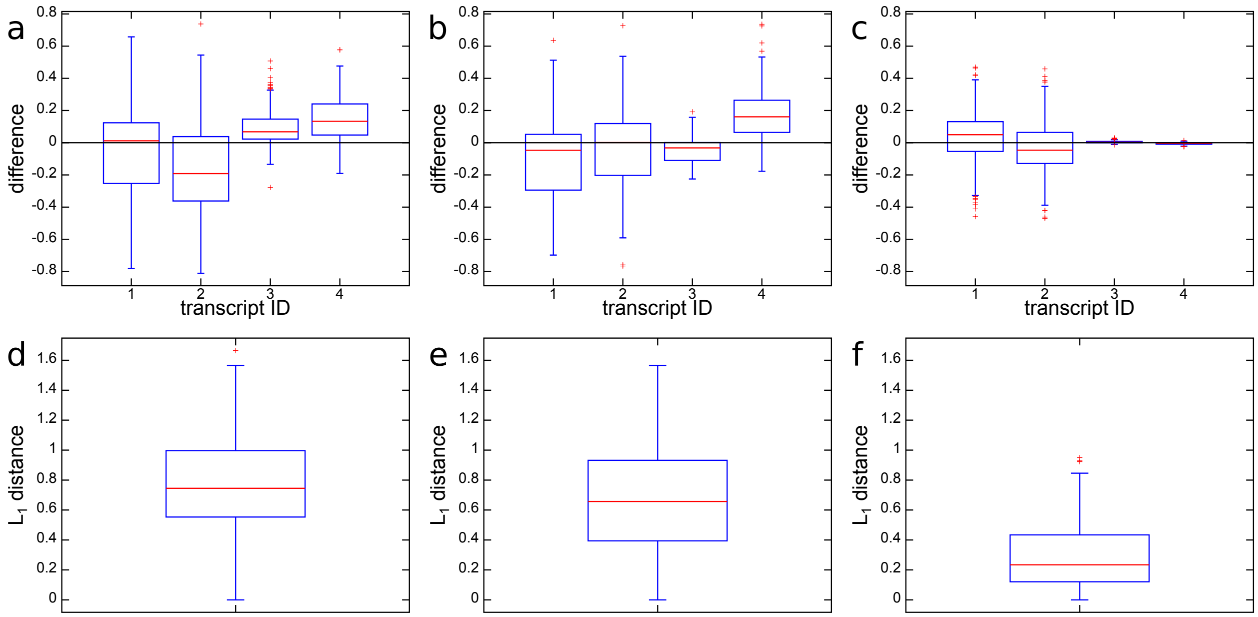

Figure 44: Difference between true and estimated abundances for **data with 5' bias and incorrect annotations**. Figures (a), (b) and (c) show boxplots for the difference for each of the 4 transcripts in the **LDHD** gene for **Cufflinks 2.2.0** (a), **PennSeq** (b) and the **4p Mix^2 model with group tying** (c). Figures (d), (e) and (f) show boxplots of the L_1 distance for Cufflinks 2.2.0 (d), PennSeq (e) and the 4p Mix^2 model with group tying (f). The mean and the standard deviation of the values in Figures (c) and (d) are given in Table 15.

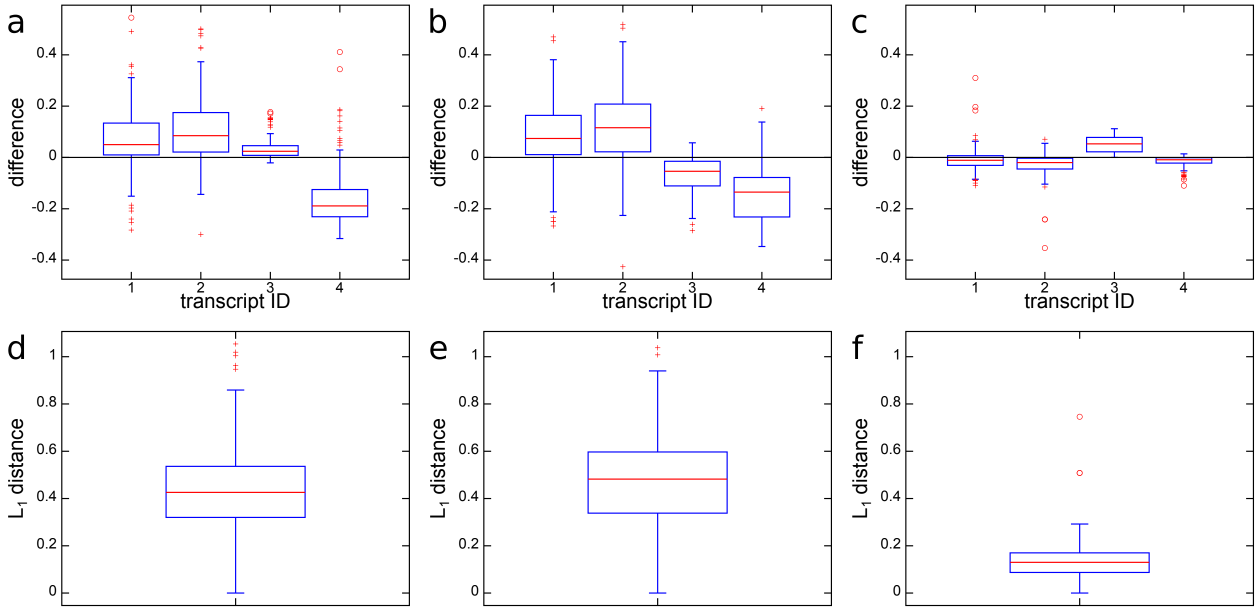

Figure 45: Difference between true and estimated abundances for **data with 3' bias and incorrect annotations**. Figures (a), (b) and (c) show boxplots for the difference for each of the 4 transcripts in the **LDHD** gene for **Cufflinks 2.2.0** (a), **PennSeq** (b) and the **4p Mix^2 model with group tying** (c). Figures (d), (e) and (f) show boxplots of the L_1 distance for Cufflinks 2.2.0 (d), PennSeq (e) and the 4p Mix^2 model with group tying (f). The mean and the standard deviation of the values in Figures (c) and (d) are given in Table 16.

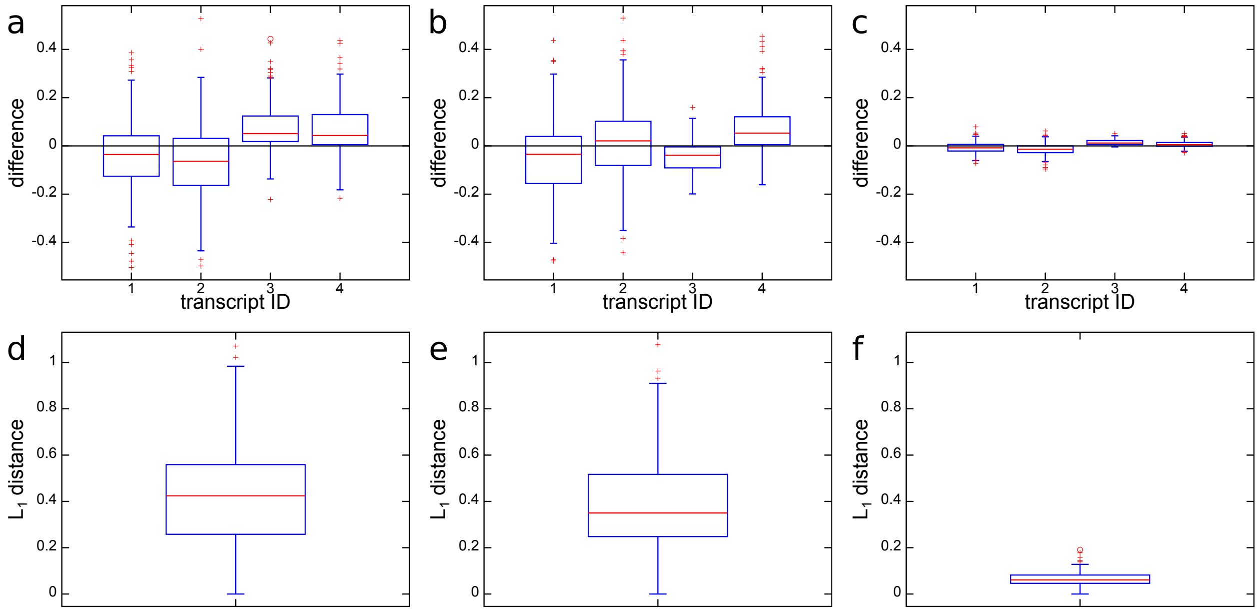

Figure 46: Difference between true and estimated abundances for **data with 5'+3' bias and incorrect annotations**. Figures (a), (b) and (c) show boxplots for the difference for each of the 4 transcripts in the **LDHD** gene for **Cufflinks 2.2.0** (a), **PennSeq** (b) and the **4p Mix^2 model with group tying** (c). Figures (d), (e) and (f) show boxplots of the L$_1$ distance for Cufflinks 2.2.0 (d), PennSeq (e) and the 4p Mix^2 model with group tying (f). The mean and the standard deviation of the values in Figures (c) and (d) are given in Table 17.

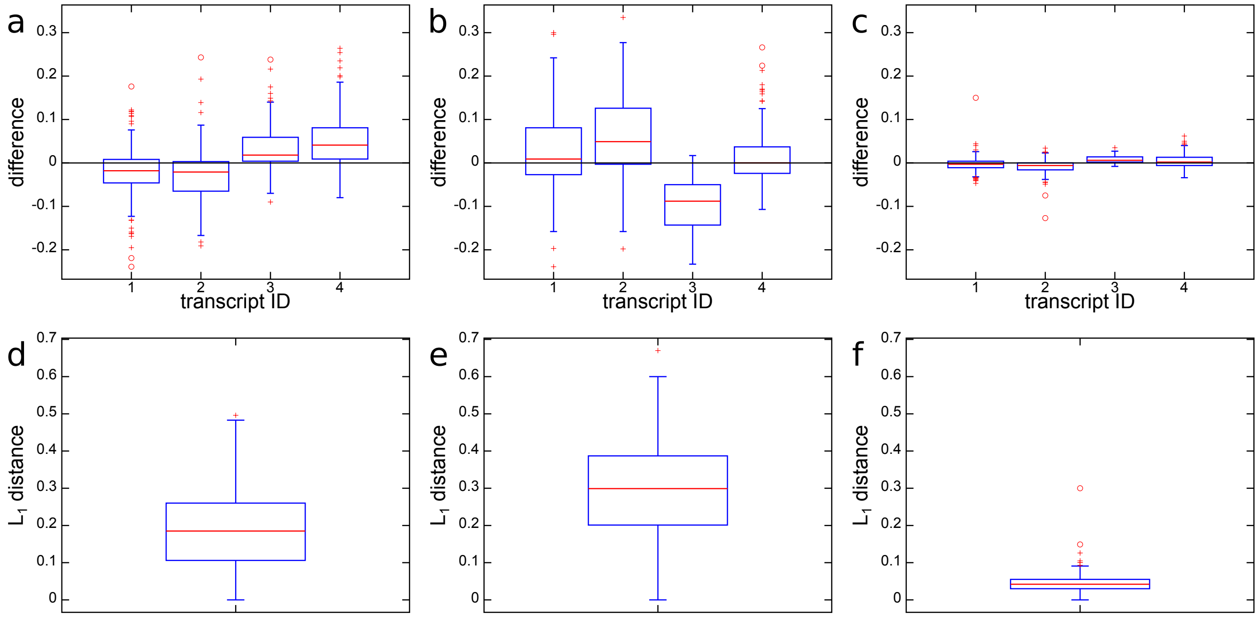

Figure 47: Difference between true and estimated abundances for **data with Cufflinks bias and incorrect annotations**. Figures (a), (b) and (c) show boxplots for the difference for each of the 4 transcripts in the **LDHD** gene for **Cufflinks 2.2.0** (a), **PennSeq** (b) and the **4p Mix^2 model with group tying** (c). Figures (d), (e) and (f) show boxplots of the L$_1$ distance for Cufflinks 2.2.0 (d), PennSeq (e) and the 4p Mix^2 model with group tying (f). The mean and the standard deviation of the values in Figures (c) and (d) are given in Table 18.

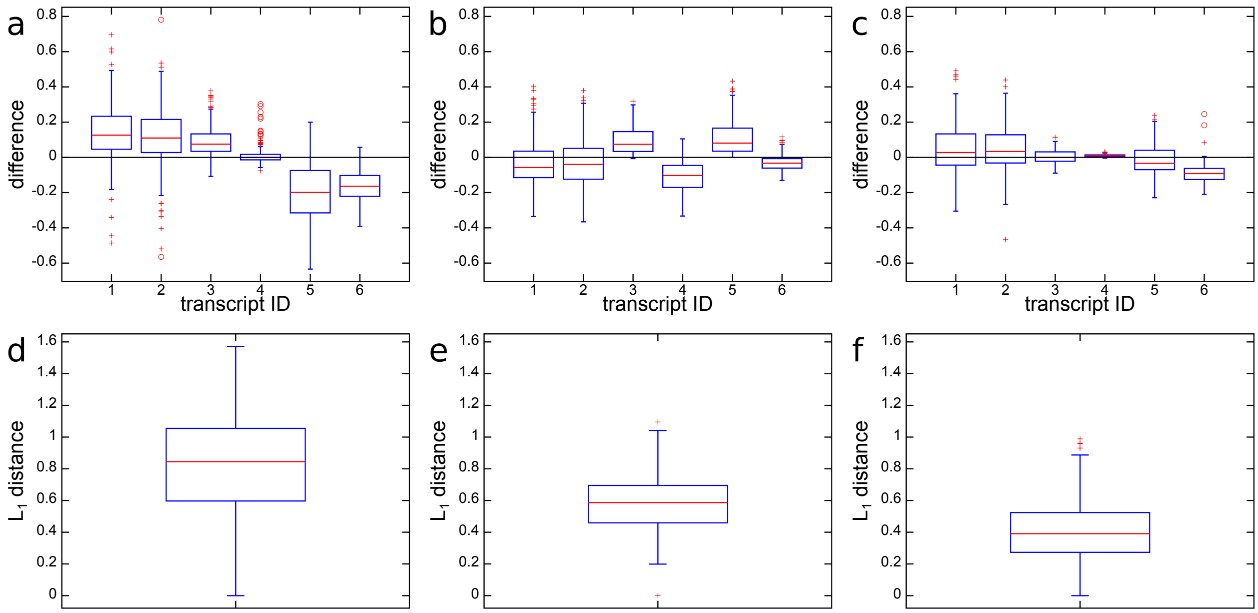

Figure 48: Difference between true and estimated abundances for **data with 5' bias and incorrect annotations**. Figures (a), (b) and (c) show boxplots for the difference for each of the 6 transcripts in the **LGALS17A** gene for **Cufflinks 2.2.0** (a), **PennSeq** (b) and the **4p Mix^2 model with group tying** (c). Figures (d), (e) and (f) show boxplots of the L$_1$ distance for Cufflinks 2.2.0 (d), PennSeq (e) and the 4p Mix^2 model with group tying (f). The mean and the standard deviation of the values in Figures (c) and (d) are given in Table 15.

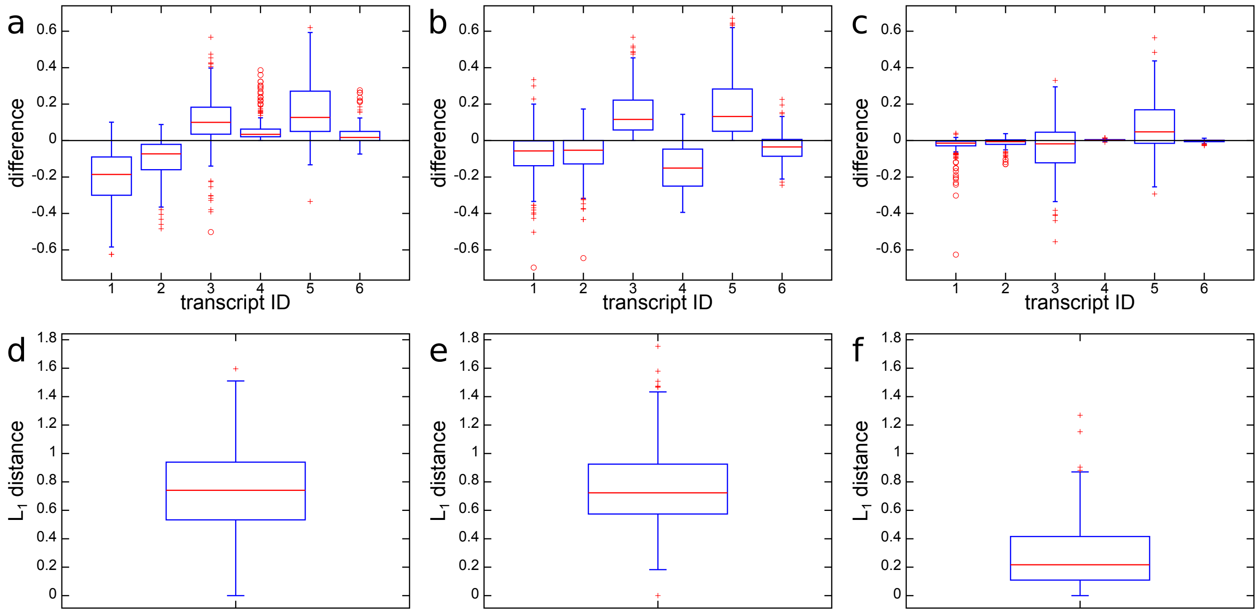

Figure 49: Difference between true and estimated abundances for **data with 3' bias and incorrect annotations**. Figures (a), (b) and (c) show boxplots for the difference for each of the 6 transcripts in the **LGALS17A** gene for **Cufflinks 2.2.0** (a), **PennSeq** (b) and the **4p Mix^2 model with group tying** (c). Figures (d), (e) and (f) show boxplots of the L$_1$ distance for Cufflinks 2.2.0 (d), PennSeq (e) and the 4p Mix^2 model with group tying (f). The mean and the standard deviation of the values in Figures (c) and (d) are given in Table 16.

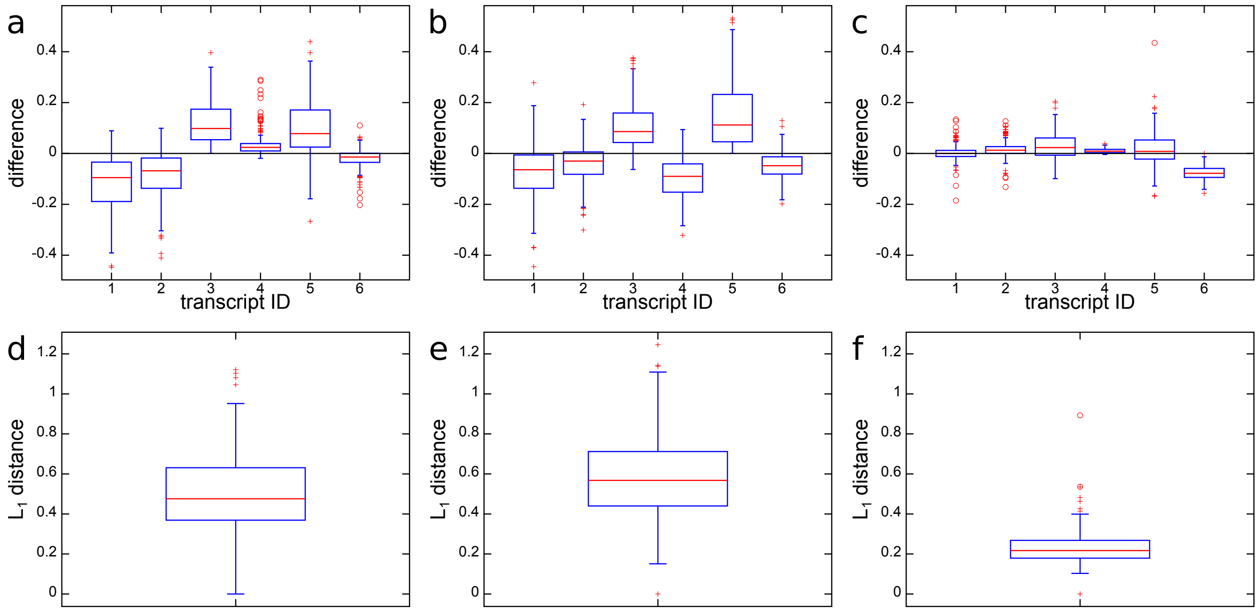

Figure 50: Difference between true and estimated abundances for **data with 5'+3' bias and incorrect annotations**. Figures (a), (b) and (c) show boxplots for the difference for each of the 6 transcripts in the **LGALS17A** gene for **Cufflinks 2.2.0** (a), **PennSeq** (b) and the **4p Mix^2 model with group tying** (c). Figures (d), (e) and (f) show boxplots of the L_1 distance for Cufflinks 2.2.0 (d), PennSeq (e) and the 4p Mix^2 model with group tying (f). The mean and the standard deviation of the values in Figures (c) and (d) are given in Table 17.

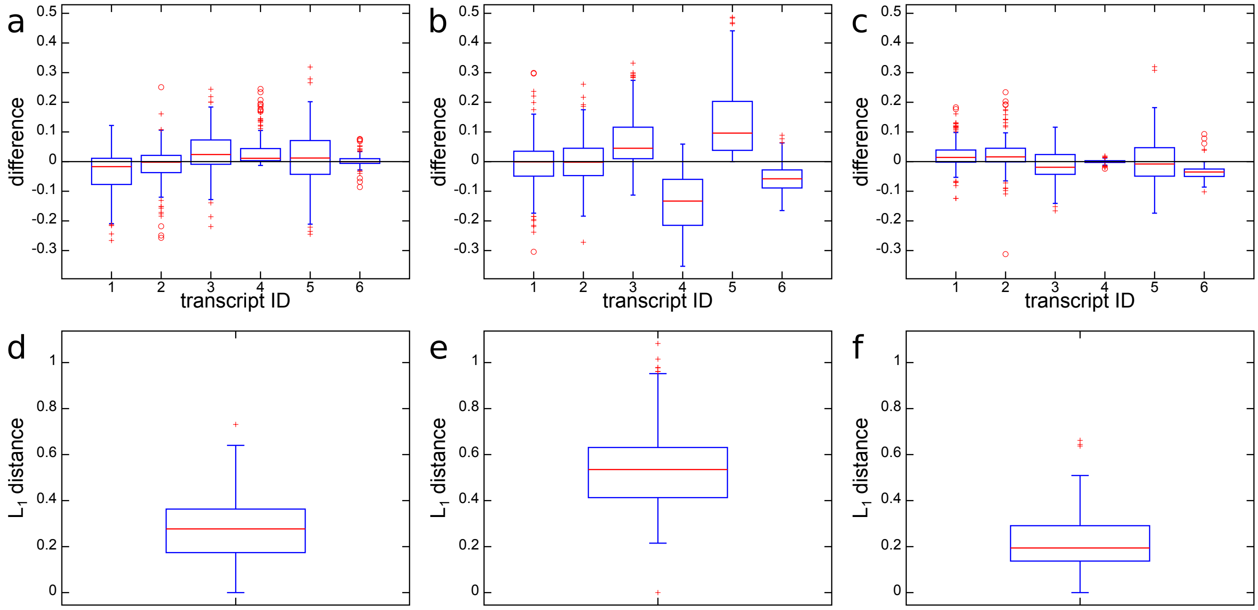

Figure 51: Difference between true and estimated abundances for **data with Cufflinks bias and incorrect annotations**. Figures (a), (b) and (c) show boxplots for the difference for each of the 6 transcripts in the **LGALS17A** gene for **Cufflinks 2.2.0** (a), **PennSeq** (b) and the **4p Mix^2 model with group tying** (c). Figures (d), (e) and (f) show boxplots of the L_1 distance for Cufflinks 2.2.0 (d), PennSeq (e) and the 4p Mix^2 model with group tying (f). The mean and the standard deviation of the values in Figures (c) and (d) are given in Table 18.

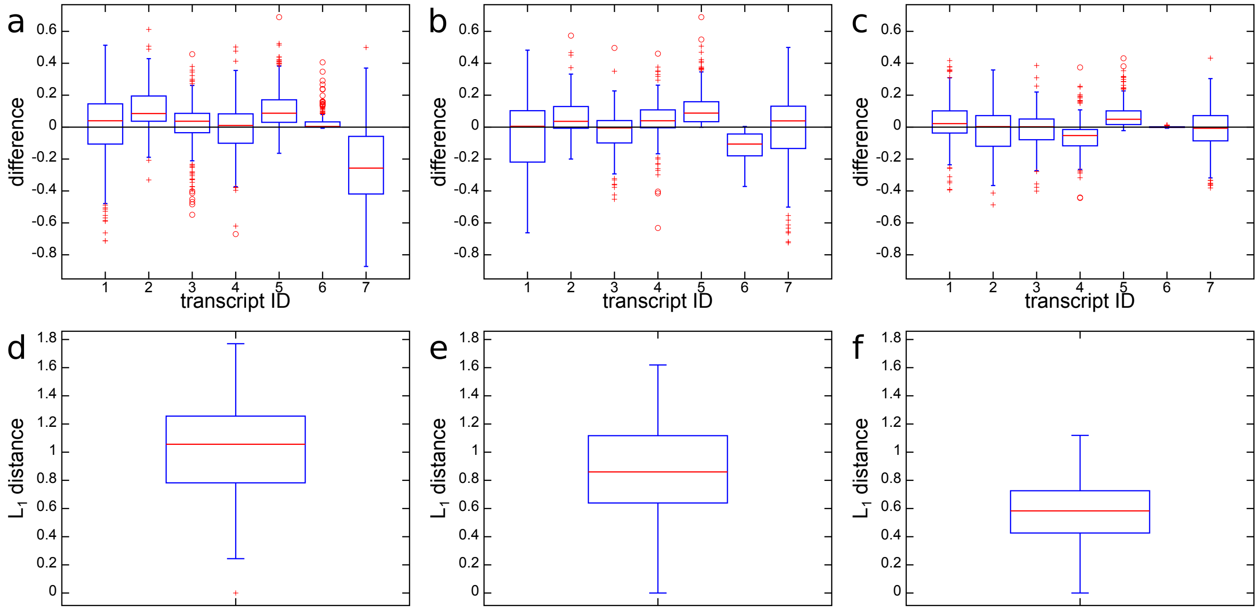

Figure 52: Difference between true and estimated abundances for **data with 5' bias and incorrect annotations**. Figures (a), (b) and (c) show boxplots for the difference for each of the 7 transcripts in the **DAPK3** gene for **Cufflinks 2.2.0** (a), **PennSeq** (b) and the **4p Mix^2 model with group tying** (c). Figures (d), (e) and (f) show boxplots of the L_1 distance for Cufflinks 2.2.0 (d), PennSeq (e) and the 4p Mix^2 model with group tying (f). The mean and the standard deviation of the values in Figures (c) and (d) are given in Table 15.

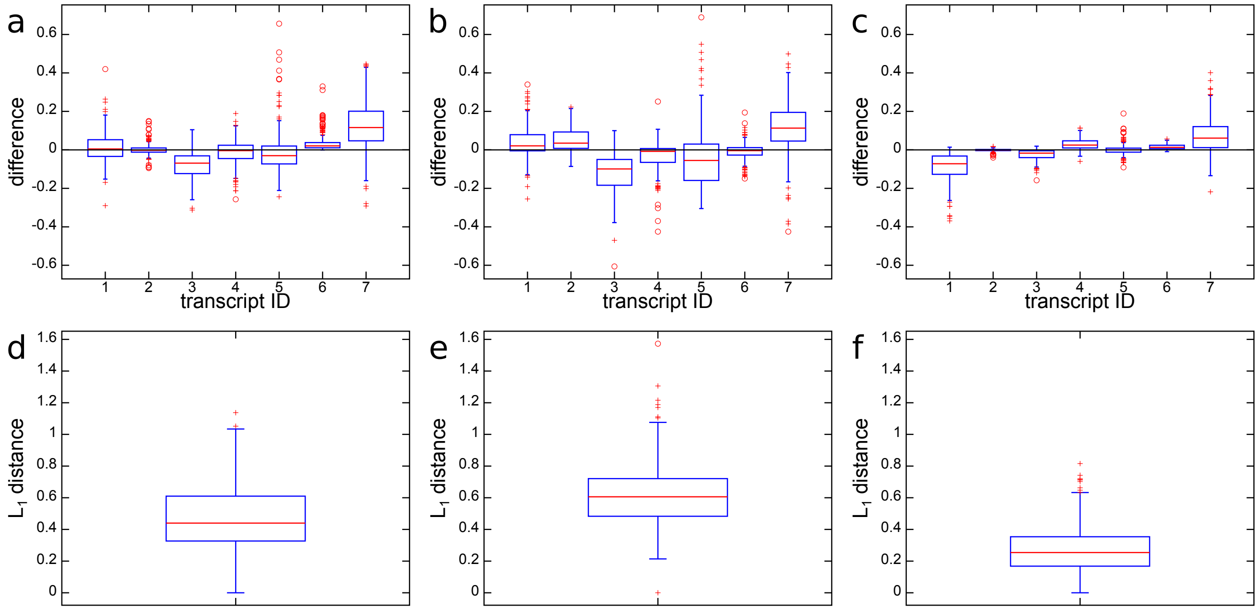

Figure 53: Difference between true and estimated abundances for **data with 3' bias and incorrect annotations**. Figures (a), (b) and (c) show boxplots for the difference for each of the 7 transcripts in the **DAPK3** gene for **Cufflinks 2.2.0** (a), **PennSeq** (b) and the **4p Mix^2 model with group tying** (c). Figures (d), (e) and (f) show boxplots of the L_1 distance for Cufflinks 2.2.0 (d), PennSeq (e) and the 4p Mix^2 model with group tying (f). The mean and the standard deviation of the values in Figures (c) and (d) are given in Table 16.

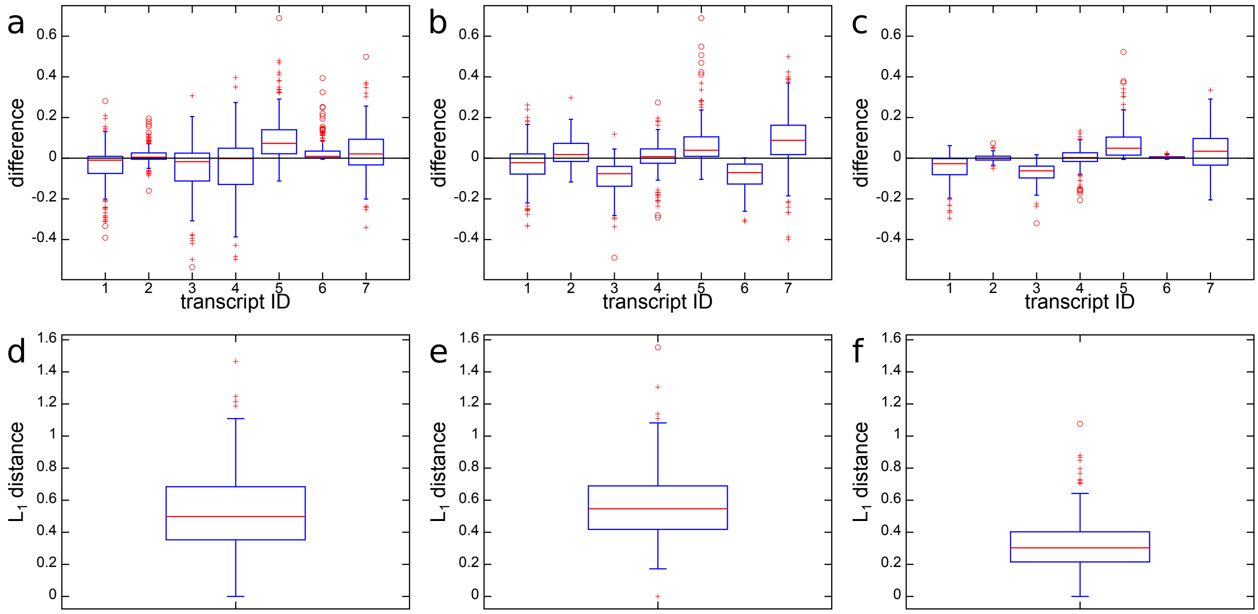

Figure 54: Difference between true and estimated abundances for **data with 5'+3' bias and incorrect annotations**. Figures (a), (b) and (c) show boxplots for the difference for each of the 7 transcripts in the **DAPK3** gene for **Cufflinks 2.2.0** (a), **PennSeq** (b) and the **4p Mix^2 model with group tying** (c). Figures (d), (e) and (f) show boxplots of the L$_1$ distance for Cufflinks 2.2.0 (d), PennSeq (e) and the 4p Mix^2 model with group tying (f). The mean and the standard deviation of the values in Figures (c) and (d) are given in Table 17.

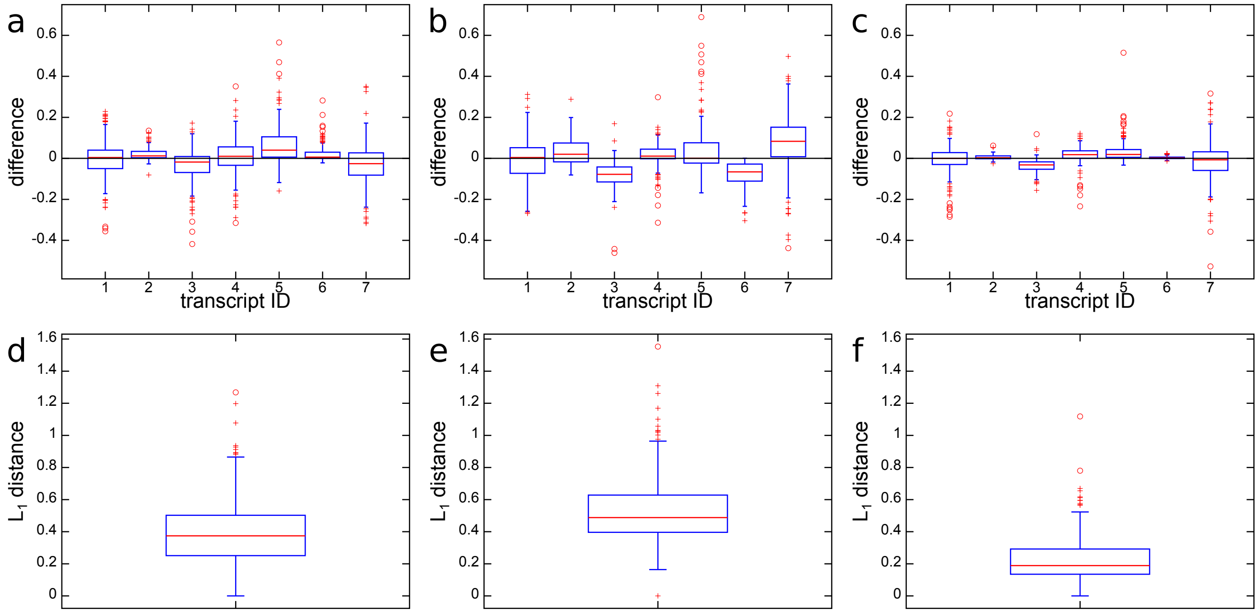

Figure 55: Difference between true and estimated abundances for **data with Cufflinks bias and incorrect annotations**. Figures (a), (b) and (c) show boxplots for the difference for each of the 7 transcripts in the **DAPK3** gene for **Cufflinks 2.2.0** (a), **PennSeq** (b) and the **4p Mix^2 model with group tying** (c). Figures (d), (e) and (f) show boxplots of the L$_1$ distance for Cufflinks 2.2.0 (d), PennSeq (e) and the 4p Mix^2 model with group tying (f). The mean and the standard deviation of the values in Figures (c) and (d) are given in Table 18.

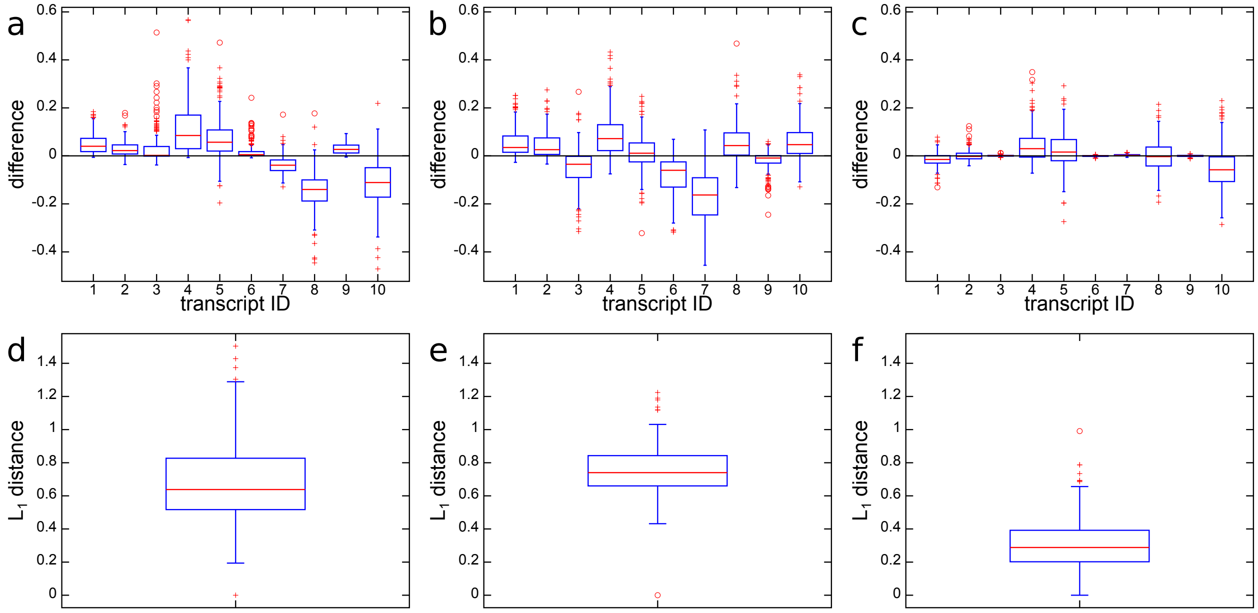

Figure 56: Difference between true and estimated abundances for **data with 5' bias and incorrect annotations**. Figures (a), (b) and (c) show boxplots for the difference for each of the 10 transcripts in the **HAUS5** gene for **Cufflinks 2.2.0** (a), **PennSeq** (b) and the **4p Mix^2 model with group tying** (c). Figures (d), (e) and (f) show boxplots of the L$_1$ distance for Cufflinks 2.2.0 (d), PennSeq (e) and the 4p Mix^2 model with group tying (f). The mean and the standard deviation of the values in Figures (c) and (d) are given in Table 15.

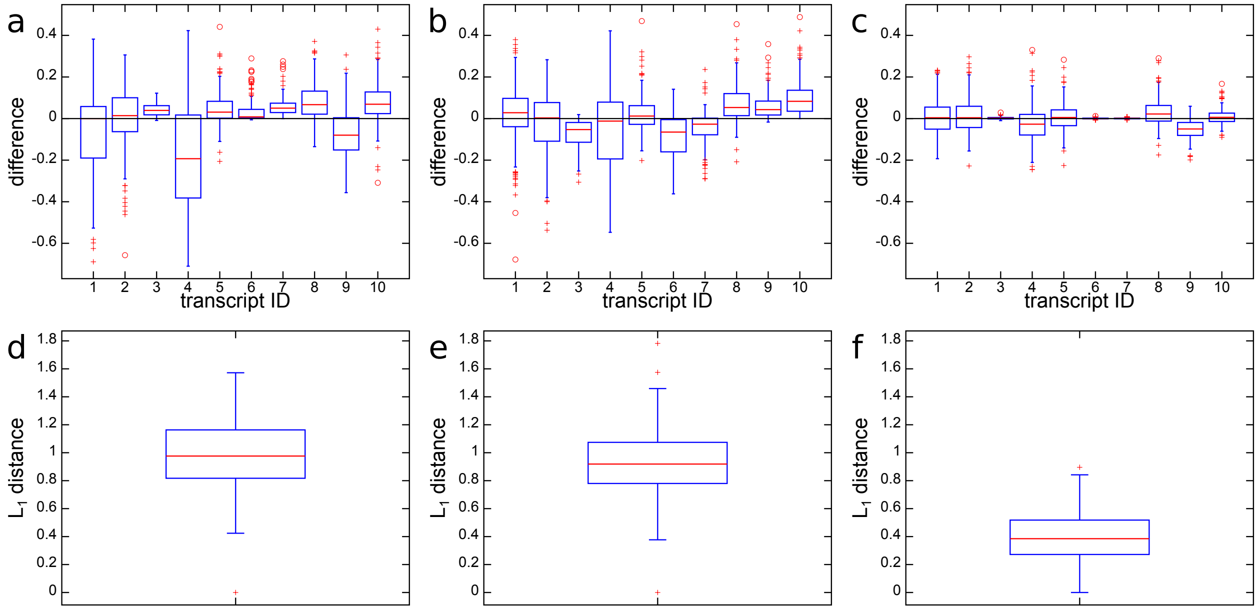

Figure 57: Difference between true and estimated abundances for **data with 3' bias and incorrect annotations**. Figures (a), (b) and (c) show boxplots for the difference for each of the 10 transcripts in the **HAUS5** gene for **Cufflinks 2.2.0** (a), **PennSeq** (b) and the **4p Mix^2 model with group tying** (c). Figures (d), (e) and (f) show boxplots of the L$_1$ distance for Cufflinks 2.2.0 (d), PennSeq (e) and the 4p Mix^2 model with group tying (f). The mean and the standard deviation of the values in Figures (c) and (d) are given in Table 16.

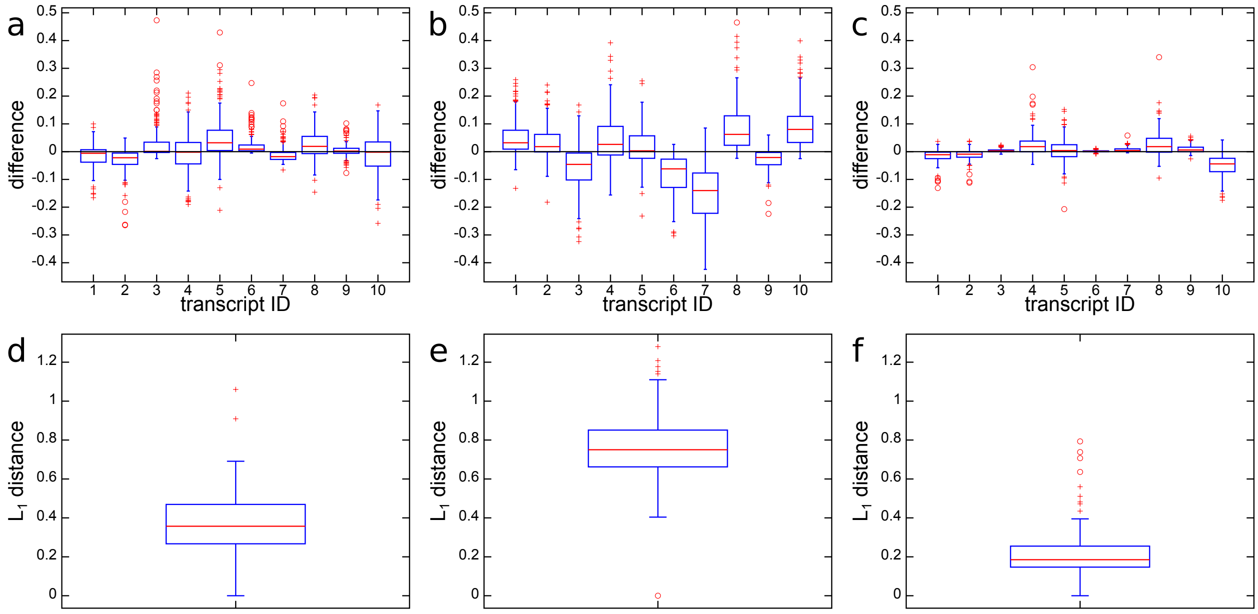

Figure 58: Difference between true and estimated abundances for **data with 5'+3' bias and incorrect annotations**. Figures (a), (b) and (c) show boxplots for the difference for each of the 10 transcripts in the **HAUS5** gene for **Cufflinks 2.2.0** (a), **PennSeq** (b) and the **4p Mix^2 model with group tying** (c). Figures (d), (e) and (f) show boxplots of the L_1 distance for Cufflinks 2.2.0 (d), PennSeq (e) and the 4p Mix^2 model with group tying (f). The mean and the standard deviation of the values in Figures (c) and (d) are given in Table 17.

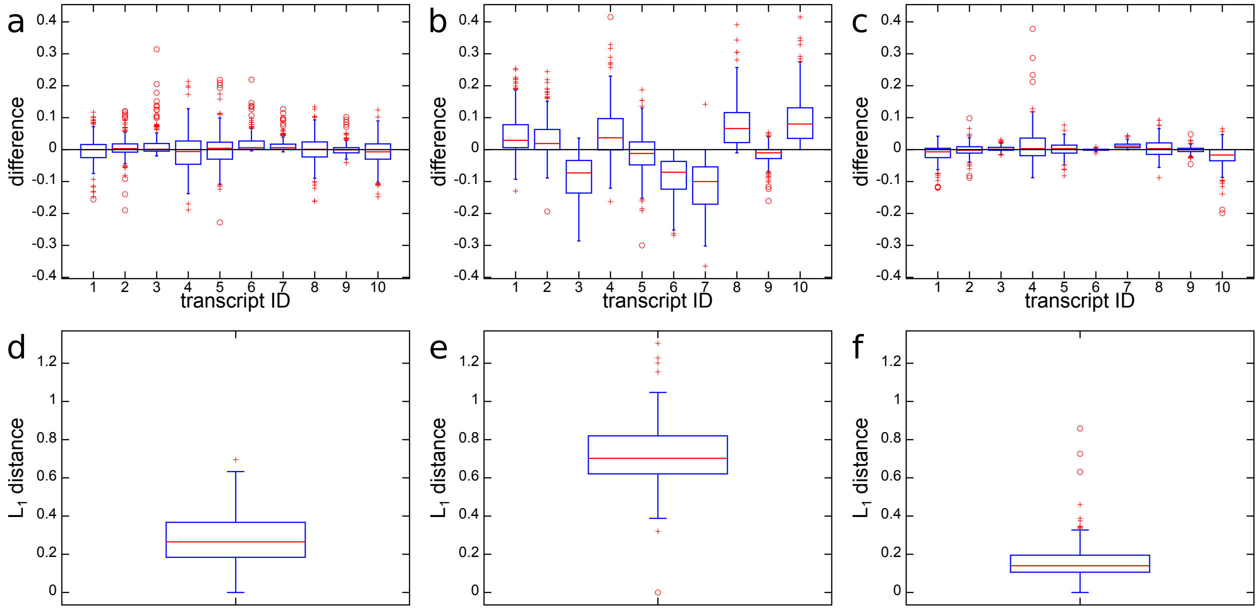

Figure 59: Difference between true and estimated abundances for **data with Cufflinks bias and incorrect annotations**. Figures (a), (b) and (c) show boxplots for the difference for each of the 10 transcripts in the **HAUS5** gene for **Cufflinks 2.2.0** (a), **PennSeq** (b) and the **4p Mix^2 model with group tying** (c). Figures (d), (e) and (f) show boxplots of the L_1 distance for Cufflinks 2.2.0 (d), PennSeq (e) and the 4p Mix^2 model with group tying (f). The mean and the standard deviation of the values in Figures (c) and (d) are given in Table 18.

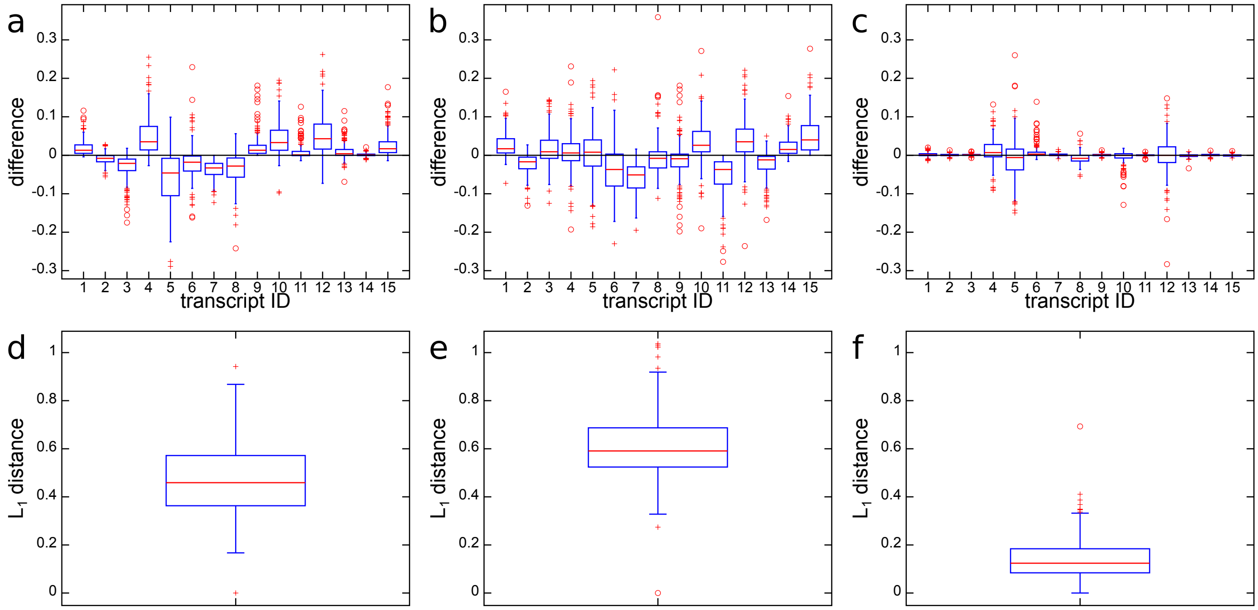

Figure 60: Difference between true and estimated abundances for **data with 5' bias and incorrect annotations**. Figures (a), (b) and (c) show boxplots for the difference for each of the 15 transcripts in the **USF2** gene for **Cufflinks 2.2.0** (a), **PennSeq** (b) and the **4p Mix^2 model with group tying** (c). Figures (d), (e) and (f) show boxplots of the L_1 distance for Cufflinks 2.2.0 (d), PennSeq (e) and the 4p Mix^2 model with group tying (f). The mean and the standard deviation of the values in Figures (c) and (d) are given in Table 15.

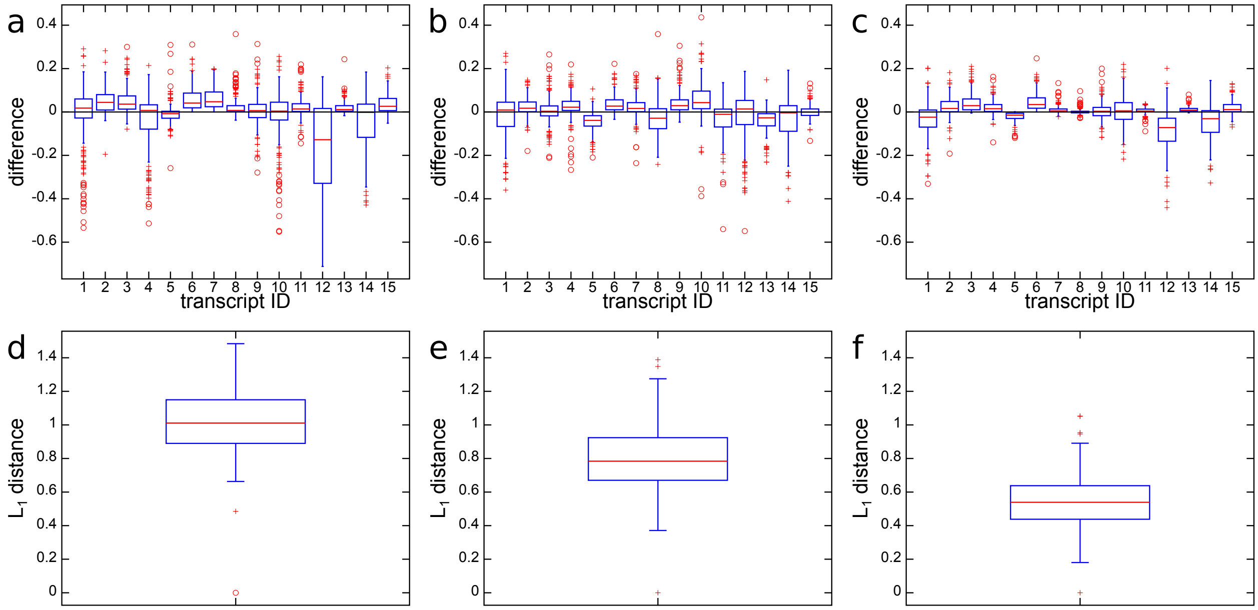

Figure 61: Difference between true and estimated abundances for **data with 3' bias and incorrect annotations**. Figures (a), (b) and (c) show boxplots for the difference for each of the 15 transcripts in the **USF2** gene for **Cufflinks 2.2.0** (a), **PennSeq** (b) and the **4p Mix^2 model with group tying** (c). Figures (d), (e) and (f) show boxplots of the L_1 distance for Cufflinks 2.2.0 (d), PennSeq (e) and the 4p Mix^2 model with group tying (f). The mean and the standard deviation of the values in Figures (c) and (d) are given in Table 16.

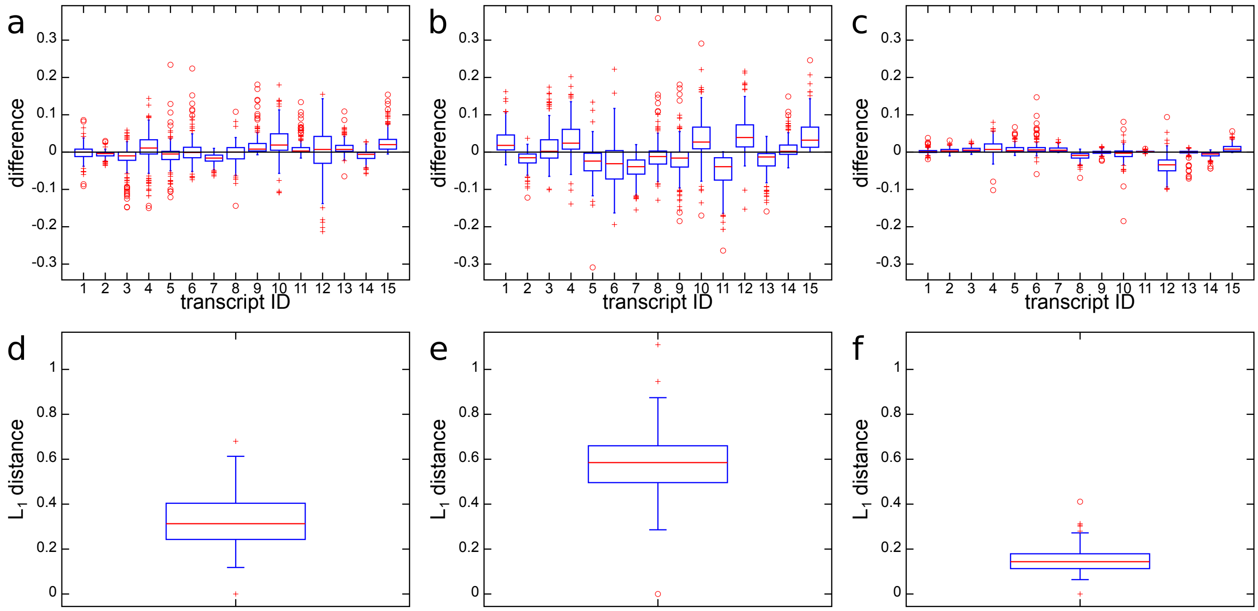

Figure 62: Difference between true and estimated abundances for **data with 5'+3' bias and incorrect annotations**. Figures (a), (b) and (c) show boxplots for the difference for each of the 15 transcripts in the **USF2** gene for **Cufflinks 2.2.0** (a), **PennSeq** (b) and the **4p Mix^2 model with group tying** (c). Figures (d), (e) and (f) show boxplots of the L$_1$ distance for Cufflinks 2.2.0 (d), PennSeq (e) and the 4p Mix^2 model with group tying (f). The mean and the standard deviation of the values in Figures (c) and (d) are given in Table 17.

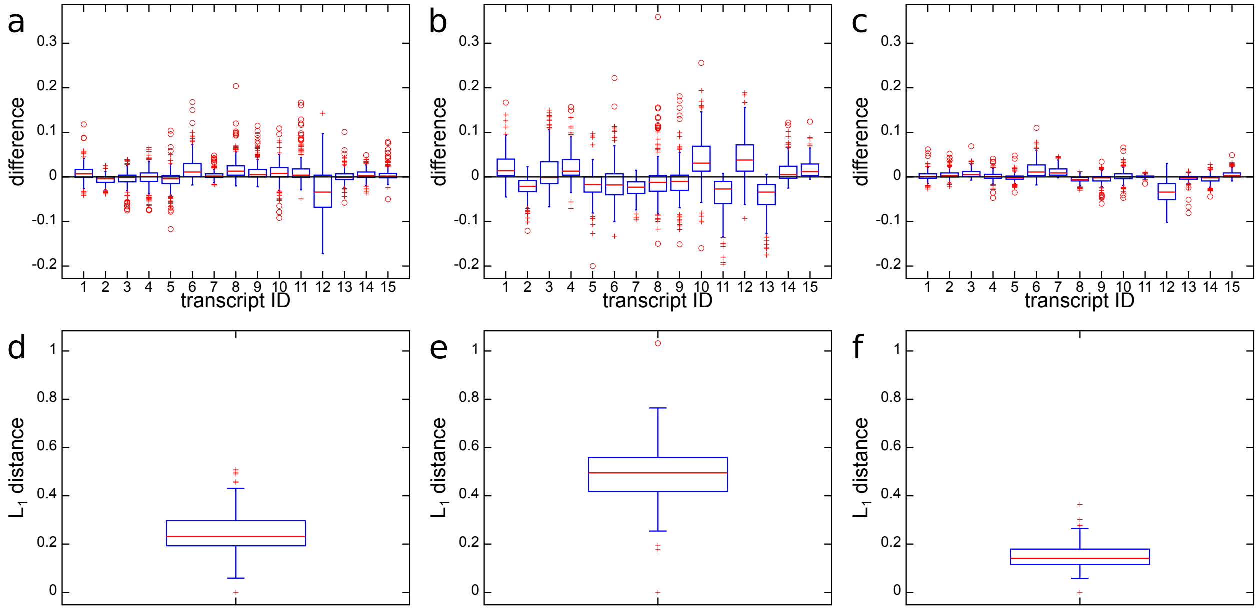

Figure 63: Difference between true and estimated abundances for **data with Cufflinks bias and incorrect annotations**. Figures (a), (b) and (c) show boxplots for the difference for each of the 15 transcripts in the **USF2** gene for **Cufflinks 2.2.0** (a), **PennSeq** (b) and the **4p Mix^2 model with group tying** (c). Figures (d), (e) and (f) show boxplots of the L$_1$ distance for Cufflinks 2.2.0 (d), PennSeq (e) and the 4p Mix^2 model with group tying (f). The mean and the standard deviation of the values in Figures (c) and (d) are given in Table 18.

References

[1] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38, 1977.

[2] Adam Roberts, Cole Trapnell, Julie Donaghey, John L Rinn, and Lior Pachter. Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome Biol*, 12(3):R22, Mar 2011.

[3] Helga Thorvaldsdóttir, James T. Robinson, and Jill P. Mesirov. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics*, 14(2):178–192, 2013.